
Photonics for Neuromorphic Computing: Fundamentals, Devices, and Opportunities

Renjie Li¹Yuanhao Gong¹Hai Huang¹Yuze Zhou¹Sixuan Mao¹Zhijian Wei^{*, 2} Zhaoyu Zhang^{*, 1}

Email: zhangzy@cuhk.edu.cn

¹ School of Science and Engineering, Guangdong Key Laboratory of Optoelectronic Materials and Chips,
Shenzhen Key Lab of Semiconductor Lasers, The Chinese University of Hong Kong, Shenzhen, Guangdong, China² SONT Technologies Co., LTD, Shenzhen, China

* indicates corresponding authors

Abstract

In the dynamic landscape of Artificial Intelligence (AI), two notable phenomena are becoming predominant: the exponential growth of large AI model sizes and the explosion of massive amount of data. Meanwhile, scientific research such as quantum computing and protein synthesis increasingly demand higher computing capacities. Neuromorphic computing, inspired by the mechanism and functionality of human brains, uses physical artificial neurons to do computations and is drawing widespread attention. Conventional electronic computing has experienced certain difficulties, particularly concerning the latency, crosstalk, and energy consumption of digital processors. As the Moore's law approaches its terminus, there is a urgent need for alternative computing architectures that can satisfy this growing computing demand and break through the von Neumann model. Recently, the expansion of optoelectronic devices on photonic integration platforms has led to significant growth in photonic computing, where photonic integrated circuits (PICs) have enabled ultrafast artificial neural networks (ANN) with sub-nanosecond latencies, low heat dissipation, and high parallelism. Such non-von Neumann photonic computing systems hold the promise to cater to the escalating requirements of AI and scientific computing. In this review, we study recent advancements in integrated photonic neuromorphic systems, and from the perspective of materials and device engineering, we lay out the scientific and technological breakthroughs necessary to advance the state-of-the-art. In particular, we examine various technologies and devices employed in neuromorphic photonic AI accelerators, spanning from traditional optics to PICs. We evaluate the performances of different designs by energy efficiency in operations per joule (OP/J) and compute density in operations per squared millimeter per second (OP/mm²/s). Putting special emphasis on photonic components such as VCSEL lasers, optical interconnects, and frequency microcombs, we highlight the most recent breakthroughs in photonic engineering and materials science used to create advanced neuromorphic computing chips. Lastly, we recognize that existing technologies encounter obstacles in achieving photonic AI accelerators with peta-level computing speed and energy efficiency, and we also explore potential approaches in new devices, fabrication, materials, and integration to drive innovation. As the current challenges and barriers in cost, scalability, footprint, and computing capacity are resolved one-by-one, photonic neuromorphic systems are bound to co-exist with, if not replace, conventional electronic computers and transform the landscape of AI and scientific computing in the foreseeable future.

Contents

1	Introduction	3
2	Conventional Neuromorphic Computing: History and Challenges	6
2.1	Brief Overview of Conventional Computing and Its Limitations	6
2.1.1	Brief Overview of Conventional Computing	7
2.1.2	Limitations of conventional computing	8
2.2	Introduction to neuromorphic computing and its advantages	9
2.3	Challenges and limitations faced by conventional neuromorphic computing	12
3	Overview of the Fundamentals of Photonics	14
3.1	Basics of photonics and light-matter interactions	14
3.2	Comparison between photons and electrons	15
4	Photonic Components for Neuromorphic Computing	16
4.1	Key devices of photonics for neuromorphic computing	16
4.2	Optical interconnects for scalable neuromorphic systems	18
4.2.1	Advancements in Photonic Components	18
4.2.2	Evolution of network topology	19
4.2.3	System Integration and Challenges	20
4.3	Optical logic gates in neuromorphic computing	20
4.4	State-of-the-art fabrication platforms for photonic devices	21
4.4.1	Lithography	22
4.4.2	Etching	22
4.4.3	3D printing technology	23
4.4.4	Challenges and Future Directions	23
5	Recent Advances in Photonic Neuromorphic Computing	24
6	Emerging Areas in Photonics for Neuromorphic Computing	29
6.1	Emerging light sources and their potential impact	29
6.2	Emerging silicon-on-insulator (SOI) paradigm and its potential impact	31
6.3	Emerging optical encoder-ANNs and their potential impact	31
7	Remaining Challenges and Future Directions	32
7.1	Strategies for improving device performance, scalability, and reliability	32
7.2	PIC's bottlenecks, remaining challenges, and future directions of neuromorphic photonics	33
8	Conclusion	33

1 Introduction

In the dynamic landscape of Artificial Intelligence (AI) [1], two notable trends have shaped its development: the exponential growth in the number of AI model parameters and the staggering amount of data generated. These factors have propelled AI to new heights and unlocked unprecedented possibilities. The first trend revolves around the expansion of AI model parameters (Figure 1). In recent years, there has been a remarkable surge in the size and complexity of deep neural networks (DNN) (Figure 2b). Deep learning models have shown remarkable capabilities in tasks such as image classification, natural language processing, and speech recognition. The increase in the number of parameters within these large deep learning models has played a significant role in their success because by augmenting its capacity to learn intricate patterns and representations, larger models can often achieve superior performance. Large language models (LLM) [2], such as chatGPT for example, have achieved unprecedented scale and popularity since April 2023. It's reported that GPT-3 and GPT-4 consists of approximately 175 billion and 1.8 trillion parameters, respectively. These parameters serve as the variables that the model learns and adapts from training data, enabling it to generate coherent and contextually relevant text responses from prompts. This remarkable expansion has been made possible by advancements in computational hardware, such as Graphics Processing Units (GPUs) and specialized hardware accelerators like Google's Tensor Processing Unit (TPU), which enable the efficient training and inference of such complex models. Simultaneously, the second trend highlights the explosion of data generation. In the digital era, it is estimated that around 2.5 quintillion bytes (2.5 exabytes, $\text{ex}=\text{a}^{18}$) of data are created daily, and over the next three years up to 2025, global data creation is projected to accumulate to more than 180 zettabytes ($\text{zetta}=\text{a}^{21}$). From social media interactions and online transactions to sensor measurements and scientific research, this data serves as a rich resource for training and fine-tuning AI models. The availability of vast and diverse datasets has in turn revolutionized AI. Data-driven approaches, often referred to as "big data", have enabled AI models to learn from enormous amounts of information, leading to improved accuracy and robustness. Additionally, advancements in data storage, processing, and cloud technologies have made it easier to handle and analyze massive datasets efficiently. The increased capacity of large models to capture and represent complex relationships, coupled with the abundance of training data, has facilitated breakthroughs in diverse domains, including healthcare, finance, transportation, robotics, metaverse, and natural language understanding.

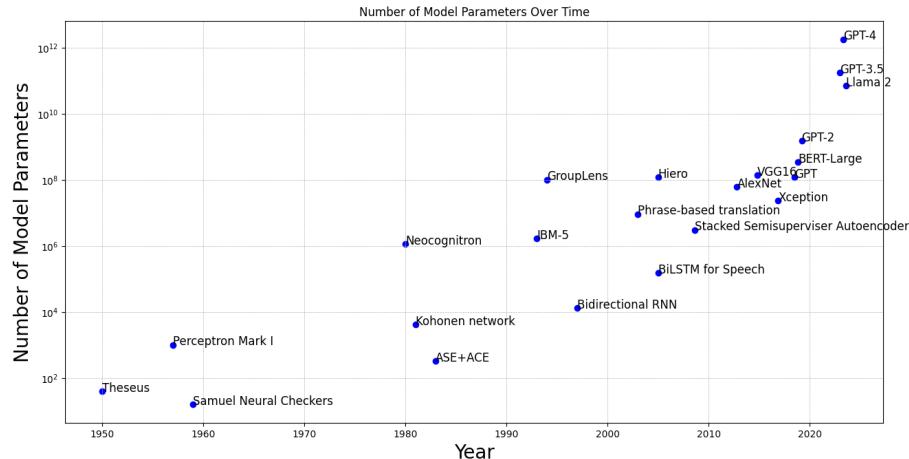


Figure 1: Trend of increase in representative AI model's parameters over time. Parameters (i.e. weights) are variables in an AI system whose values are adjusted during training to establish how input data is correlated with the output labels. The most recent chatGPT has a record-breaking 1.8 trillion parameters and is still growing. No. of parameters are estimated based on published statistics in respective papers and naturally come with some uncertainty.

However, this growth is not without challenges. The large number of model parameters and the amount of data required for training pose significant computational and resource-intensive requirements. Developing new chip hardware and computing paradigms to scale AI models efficiently, along with

addressing concerns related to latency, parallelism, and energy consumption, are areas that researchers and engineers are actively exploring.

Neuromorphic computing [3–7], an emerging field of research, focuses on developing computational systems inspired by the structural and functional characteristics of the human brain. This discipline strives to overcome the limitations of conventional computing by mimicking the parallelism, fault tolerance, and energy efficiency observed in biological neural networks (Figure 2a). At its core, neuromorphic computing aims to create hardware and software architectures that replicate the behavior of neurons and synapses. These architectures enable the processing of information using spiking neural networks and specialized neuromorphic chips, which offer real-time handling of complex data and the potential for accelerated machine learning algorithms. The primary advantage of neuromorphic computing lies in its ability to process information in a massively parallel manner, leading to enhanced computational and energy efficiency. By leveraging the inherent capabilities of neurons and synapses, neuromorphic systems demonstrate potential applications in areas such as pattern recognition, image classification, autonomous driving, and LLM. In hardware design, researchers are exploring novel components like micro-ring resonators and VCSEL lasers to emulate synaptic connections or to perform matrix operations. These photonic components enable the creation of highly efficient AI hardware that can adapt and learn from input data, similar to the plasticity observed in biological neurons. Neuromorphic techniques enable the development of self-learning systems capable of processing sensory input and making intelligent decisions in dynamic environments to facilitate the growth of AI.

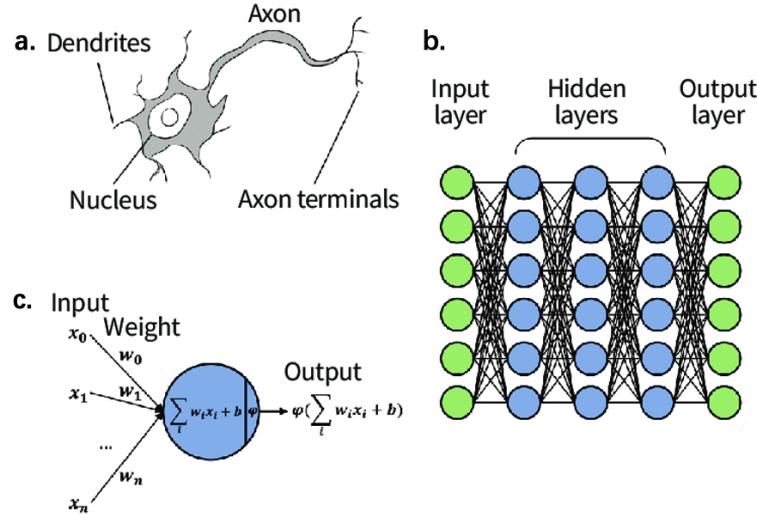


Figure 2: a. Biological neuron in animals; b. Multi-layer perception neural networks (MLP) or fully-connected (FC) layers; c. Forward propagation of artificial neurons in MLP, including the input, weighs, summation, activation function, and the output. Data obtained from online image libraries.

In optical computing, photonic devices that utilize optical near-fields and effective interactions are very important. Key components within neural networks (Figure 2b-c), such as activation functions [8–11] and backpropagation [12], currently represent the focal point of research in the realization of optical neural networks. Y. Zuo et. al. demonstrated an all-optical neural network (AONN) which can work with nonlinear activation [11]. They used spatial light modulators and Fourier lenses to realize programmed linear operations and laser-cooled atoms with electromagnetically induced transparency to realize nonlinear activation function. In this device, linear operations are done via Fourier transform and all diffracted beams in the same direction are summed onto a spot, which can be expressed as $z_i = b_i + \sum_j W_{ij} x_j$ (Figure 2c). In this formula, W is weights, x is input, and b is bias. When nonlinear operations are also done, we can get the final output $y_i = \varphi(z_i)$, where φ is a nonlinear activation function (Figure 2c). It is obvious that this neural network which uses Fourier lenses is easy to construct, low in power consumption, and boasts rapid computational speed. However, as a neural network system, the Fourier optical lens system might be excessively bulky.

Another all-optical neural network, termed "D2NN", is considered as a potential contender for miniaturized all-optical diffractive deep neural networks [13]. In this system, each point on a

specified layer is a neuron, and has a designed transmission or reflection coefficients. When it works, the input of each layer is defined by previous layer based on free-space diffraction. In this work, the device is on the macro-scale as it spans a few centimeters in length and width, and a few millimeters in thickness. However, this is because it operates in the terahertz range, and if altered to function in the near-infrared range suitable for optical communications, its size could potentially be reduced by a few hundred times. The fabrication of such tiny devices currently presents a significant challenge, yet D2NN remains a viable contender for integrated AONN systems.

With its rapid development, relative maturity, and high degree of integration, silicon photonics emerges as a compelling candidate for the integration of AONN. In silicon photonic integrated circuit, the Mach-Zehnder Interferometer (MZI) can regulate the intensity of optical signals through light interference. By amalgamating multiple MZIs, we can form an integrated optical neural network (ONN) on-chip. In 2017, Y. Shen et. al. proposed an AONN based on cascaded array of 56 programmable MZI in a silicon photonic integrated circuit [14]. By programming the internal and external phase shifters of each MZI, this system enables arbitrary SU(4) rotations and finally expresses unitary matrices. The splitting ratio and the differential output phase are controlled when the operations between unitary matrices are done. In 2019, I. Williamson et. al. added nonlinear activation function operating on MZI phase shift and realized nonlinear activation function on on-chip AONN [8]. They improved test accuracy on the MNIST task from 85% to 94%.

With earlier development of nonlinear activation functions, AONNs, and D2NNs that serve as the building blocks of photonic neuromorphic computing, researchers later began to explore photonic devices such as frequency microcombs [15], micro-ring resonators [16], lasers [17], metasurfaces [18], optical attenuators [19], and photodiodes [19] etc. to realize more advanced neuromorphic functionalities. We can roughly categorize existing efforts into three branches: 1. simulate the principles of forward propagation in an artificial neuron; 2. achieve image classification or pattern recognition; 3. realize convolutional operations by performing matrix-vector multiplications (MVM). A full timeline of the evolution of photonic neuromorphic systems is shown in Figure 3.

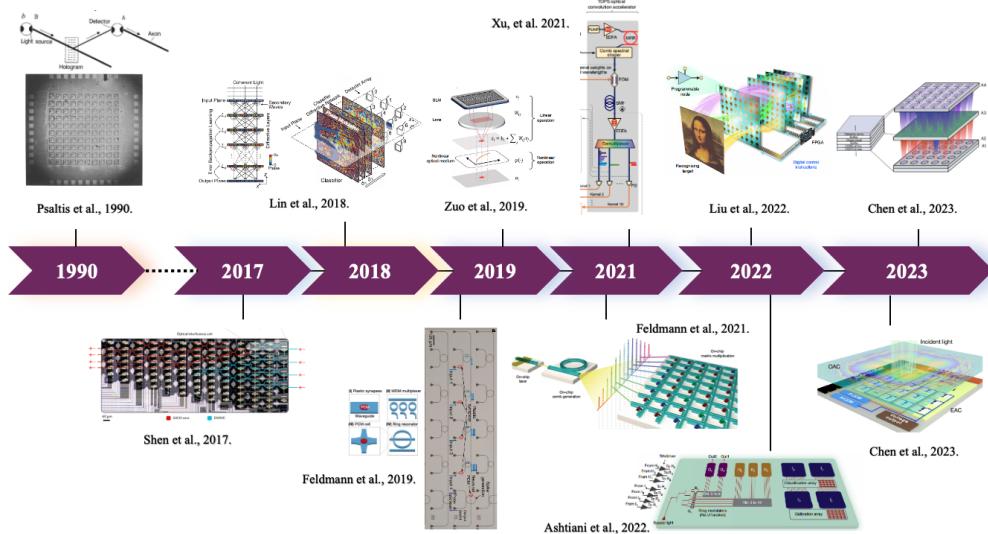


Figure 3: Timeline of the evolution of photonic neuromorphic systems, especially for AI applications. Existing works can be categorized into three branches: 1. simulate the principles of forward propagation in an artificial neuron; 2. achieve image classification or pattern recognition; 3. realize convolutional operations by performing MVM. Sources are from milestone works spanning from the 90s to 2023: [11, 13–16, 19–21] [17, 18, 22]. For detailed specs of each work, refer to Table 1.

About a dozen reviews have been written on the topic of photonic neuromorphic computing systems. Some place heavier emphasis on a specific aspect of the system such as spike generation, synapses or nonlinearity, and some focus on specific structures such as MZI, frequency combs, or diffractive layers, and some others tend to discuss the relationship between photonic neural networks and machine learning via the bridge of multiply-accumulate operations (MAC) and wavelength division

multiplexing (WDM). This review, however, aims to concisely and comprehensively unify each of the aforementioned aspects of photonic neuromorphic design and dissect them from the perspective of photonic engineering and materials science; as a result, details/fundamentals of the theories of neural networks or computing systems will not be fully addressed here. According to Figure 4, the review is organized as follows: Sec. 1 Introduction. Sec. 2 Outlines the history and challenges of neuromorphic computing based on conventional electronics paradigm. Sec. 3 Introduces fundamentals of photonics, including theories and principles of light-matter interaction. Sec. 4 Covers existing photonic components and fab platforms for neuromorphic computing. Sec. 5 Details recent advances of photonic neuromorphic computing, providing a summary of the evolution of photonic neuromorphic paradigms (also graphed in Figure 3). Sec. 6 Addresses emerging technologies in photonics for neuromorphic computing, including topological insulators and PCSEL lasers. Sec. 7 Discusses the challenges and future directions in photonic neuromorphic computing. Finally, Sec. 8 summarizes the review by providing a conclusion and closing marks.

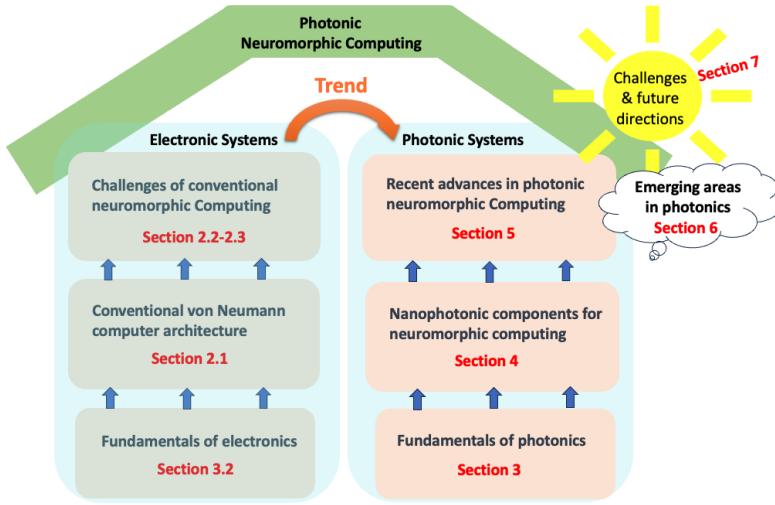


Figure 4: Overall hierarchical organization of this paper and logical links between all sections.

2 Conventional Neuromorphic Computing: History and Challenges

2.1 Brief Overview of Conventional Computing and Its Limitations

With the development of electronic device and improvement of the computing performance, many changes taken place in human society, and many scientific fields have made numerous breakthroughs. In order to meet the needs of scientific development, the size of electronic device is required to continue to scale to support the continuous growth of computing performance and maintain this increasing trend [23]. In particular, nowadays, because of the fast-paced advancements of AI, LLMs, Internet of Things (IoT), metaverse, and cloud computing, the amount of data has virtually exploded and the demand for high-performance memory and computing efficiency has become higher and higher [3, 24]. However, there are currently some difficulties that need to be addressed urgently. On one hand, as the Moore's law comes to its end, the required performance gains can no longer be achieved through conventional scaling device [23, 25]. On the other hand, at present, most computers that are general-purpose devices are designed based on the traditional Von Neumann architecture. Limited by the bottlenecks of hardware fabrication technology and the inherent structural problems such as "memory wall", conventional von Neumann architecture can not sustain processing large amounts of data in the AI era [4, 25]. So it is critical to propose alternative architectures that scale beyond von Neumann to break through the computing bottlenecks. Figure 5 plots figure of merit (FoM) of existing digital electronic hardware and emerging photonic neuromorphic systems as AI accelerators. We evaluate the performances of different architectures by energy efficiency in terms of operations per joule (OP/J) and compute density in terms of operations per squared millimeter per second (OP/mm²/s). By comparison, lab-tested photonic neuromorphic systems already outperform commercial electronic hardware in both energy efficiency and compute density as of February 2024.

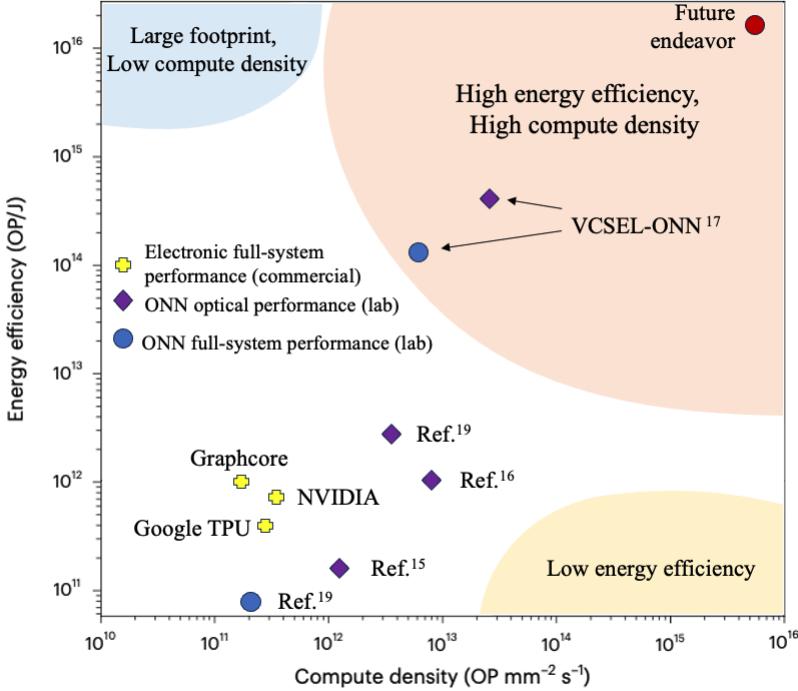


Figure 5: Comparison of state-of-the-art AI accelerators between conventional digital electronics and the emerging ONNs empowered by photonic neuromorphics. Comparison is embodied by the FoM such as energy efficiency (TOP/J) and compute density ($TOP/mm^2/s$). Corresponding FoM values of each reference can be found in Table 1. On both the x and y axis, larger values are better. As indicated by the legend, it should be noted that while the electronic systems shown here are commercial products, the photonic (ONN) systems are laboratory demonstrations. Future endeavor calls for continued efforts to improve the energy efficiency and compute density of photonic systems. Adapted with permission.[17] Copyright 2023, Springer Nature.

2.1.1 Brief Overview of Conventional Computing

Progress in the computing domain has significantly relied on semiconductor devices’ miniaturization. In 1965, Intel cofounder Gordon Moore paid attention to the stable growth rate of miniaturization and published a prediction known as “Moore’s Law”, which states that the number of transistors in each new-generation of computer chips would double every two years [26]. With the miniaturization of transistors, the density of transistors in chips increases continuously, and the computing capability is also becoming stronger and stronger. Moore’s Law has driven progress in many aspects of computing, such as better CPU performance, improved energy efficiency, larger storage capacity and better cost savings [27]. Moreover, a computer architecture named the “Von Neumann Architecture” revolutionized computing as Moore’s Law took effect, and enabled hardware engineers to build a variety of computational systems [4].

Von Neumann Architecture, also known as the Princeton Architecture, is the fundamental organizational structure of a digital computer based on the principles proposed by mathematician John von Neumann. It treats instructions (the computer program) as a special type of data and stores instructions and data in different addresses in the same memory. The main characteristics of the Von Neumann architecture is that it adopts a binary system and computations execute in a procedural order. The invention of this type of architecture laid the foundation for modern computer architectural concepts [28].

Existing digital computers are built upon the Von Neumann architecture and based on the silicon microelectronics platform. Generally, a Von Neumann computer is mainly composed of a memory bank for storing data and instructions and a central processing unit (CPU) for performing nonlinear operations and connecting transmissions between the two [6, 29, 30]. Among them, as the core of

the architecture, the CPU is composed of a control unit and an arithmetic logic unit (ALU). Usually, the CPU executes a series of instructions to process data stored in memory units through interacting with the memory system [31]. Main memory is the key to memory system and set up by adopting the Dynamic Random Access Memory (DRAM) technology. When performing computing, data in main memory needs to be processed. Because the CPU can only process the data in the cache, memory controller will send a series of instructions to the DRAM module through the off-chip bus, receive the response from the DRAM module after receiving the instructions, and then use the cache or registers to save the data [32].

2.1.2 Limitations of conventional computing

Continuous expansion in the amount of computations and data leads to stricter requirements for high-performance, and computer architecture needs to shift from the intensive type of computing to the intensive type of memory [3]. Nowadays, because of many limitations, it is hard for traditional computing to adapt to the current development pace.

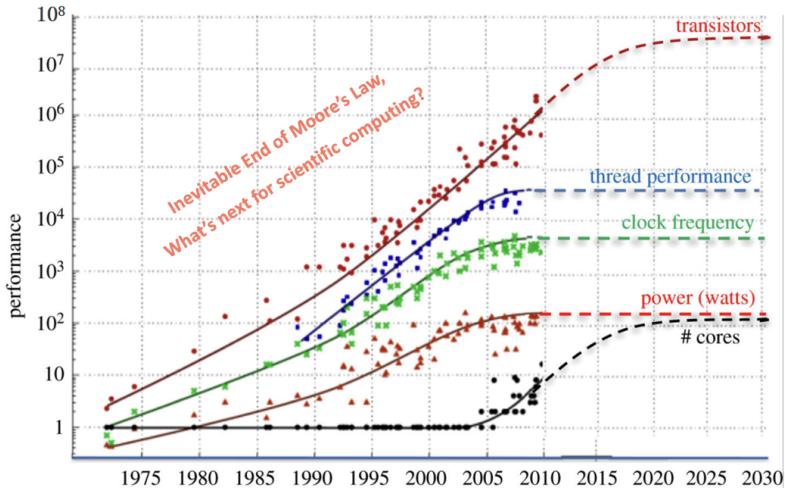


Figure 6: Moore’s Law timeline, including Moore’s Bend with transistors/CPU inflected with multi-Core CPUs beginning in 2005. Improvement of computing performance has been greatly challenged by the end of Dennard scaling in 2004. All additional approaches to further performance improvements will end in approximately 2025 due to the end of the roadmap for improvements to semiconductor lithography. Adapted with permission.[33] Copyright 2005, ACM.

On one hand, as the scaling degree of device feature size approaching its immutable physical limit, the trend of rapid growth in aspects of speed and integration of microelectronic devices represented by complementary metal oxide semiconductors (CMOS) has slowed down [3, 34]. The 10-nm technology of Intel, originally scheduled to be launched in 2016, was delayed until 2019. Apple and Samsung, despite claiming to have reached 5-nm or even 3-nm in their nanofab, can not keep up with Moore’s Law in updating their miniaturization technology [26]. Obviously, the effects of improving traditional computational performance by reducing device size have gradually weakened, and Moore’s Law that has influenced the growth of computational capability in electronic devices is inevitably coming to an end. In other words, the Post-Moore era is coming [24] (Figure 6).

On the other hand, the Von Neumann architecture on which traditional computing is built faces problems such as “memory wall” that limits system performance, making it difficult to meet modern requirements for efficiency, energy consumption, density and cost in high-performance computing. Under the Von Neumann architecture, instructions and data are placed in the same memory, and instructions follow serial execution rules, so both of them can not be accessed simultaneously to avoid confusion in memory access [29, 34, 35]. Meanwhile, the Von Neumann architecture features a separate memory and computing architecture, where memory unit and computing unit remain separate [29]. During computing, data is frequently transferred between the memory unit and the processor unit, resulting in non-negligible delays and consuming a significant amount of energy. Up to now, under serial execution rules, data movement induces longer signal delay and increased energy

losses, because of the huge gap between the operation speed of the CPU and the speed of accessing memory (the computing speed of the CPU has far exceeded the speed of accessing memory), the restriction caused by the bandwidth of the memory hierarchies, and the heat dissipation issue caused by unresolved leakage. These conditions ultimately result in insufficient utilization of hardware resources, increased energy consumption and decreased computational efficiency. For example, the AI facial recognition network developed by Google utilized a total of 16000 CPU cores on a three-day training session while consuming 100 kilowatts of power [3, 24, 32, 34–37]. To alleviate these burdens, computing capability can be enhanced by separately increasing the bandwidth of the memory and using graphics processing unit (GPU) or AI accelerator. However, this approach has a limited benefit in terms of improving computing speed and energy consumption, and is not a long-term sustainable plan [24]. In addition, application-specific integrated circuits (ASIC) also play a role in addressing relevant issues. Compared with GPU, ASIC can significantly reduce energy consumption, but most of the consumption during operation is still wasted in the data movement rather than logic operation [38].

Overall, considering the ending of Moore's Law and the limitations of the von Neumann architecture on the further development of modern computing facilities, there is a vital need to break through the core architectural bottleneck and seek alternative architectures and paradigms to build non-von Neumann systems to prominently strengthen computing performance [36, 39].

2.2 Introduction to neuromorphic computing and its advantages

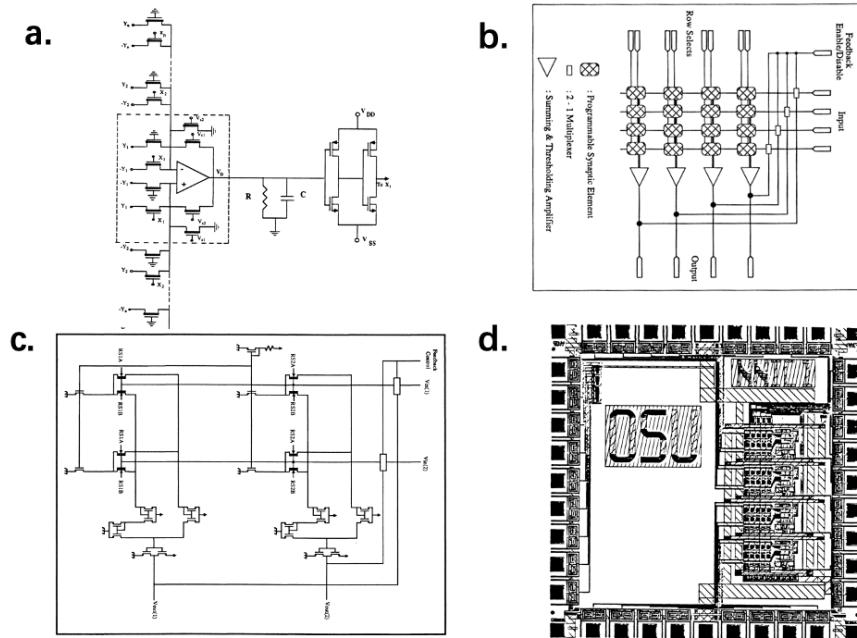


Figure 7: Analog VLSI implementation of neural systems developed on or before 1989, marking the very first attempts to construct neuromorphic chips from electronics. a. A MOS circuit for the VLSI Implementation of Hopfield-like neural networks; b. Floating gate neural network; c. Two neuron circuit with four weight synapses; d. Layout of 4 neuron chip. Reproduced with permission.[40] Copyright 1989, Springer.

Current computer technology is facing two important bottlenecks: the memory wall effect of the "von Neumann" architecture causes low energy efficiency [41–43], and Moore's Law, which leads the development of semiconductors, is expected to expire in the next few years [44–47]. On one hand, the traditional processor architecture converts the processing of high-dimensional information into a one-dimensional processing of pure time dimension, which has low efficiency and high energy consumption [48]. This architecture cannot construct appropriate algorithms when processing unstructured information, especially when processing intelligent problems in real time. In addition,

the information processing takes place in the physically separated CPU and memory. Programs and data are sequentially read from the memory into the CPU for processing, and then sent back to the memory. This process causes a large amount of energy consumption [48]. A mismatch between the rate at which programs or data are transferred back and forth and the rate at which the CPU processes information results in a severe memory wall effect [41, 49, 50]. On the other hand, as the semiconductor industry enters the sub-10 nm threshold, devices are approaching the limits of their physical shrinkage, and quantum effects are increasingly interfering with the normal operation of electronic devices [51, 52]. Although people have different estimates of the specific end time of Moore's Law, there is no controversy in the industry about the end of Moore's Law that has lasted for the past 50 years.

In 1989, Caltech's Carver Mead proposed the concept of "neuromorphic engineering (or brain-like computing)" in his book titled "Analog VLSI implementation of neural systems", which uses sub-threshold analog circuits to simulate Spiking Neural Network (SNN) [40]. As shown in Figure 7, Mead showed how the powerful organizing principles of animals' nervous systems can be realized in silicon integrated circuits, where examples include silicon neural systems that replicate animal senses. Meanwhile, Moore's Law continued to develop, and the frequency and performance of processors based on the von Neumann architecture continued to improve, while brain-inspired computing remained stagnant for more than 10 years. Around 2004, the main frequency of single-core processors stopped growing, and while the industry turned to multi-core processors, the academic community began to seek alternative technologies such as non-von Neumann architectures. Since that point and in the following 10 years or so, neuromorphic computing has begun to attract widespread attention.

It is well known that the nervous system of mammals, especially humans, is one of the most efficient and robust structures in nature. The human brain has a large number of connections and exhibits strong parallelism. It has about 10^{11} neurons and 10^{15} synapses, but consumes only about 20W of energy [59]. Neurons achieve biological interconnection at a speed of a few milliseconds and have excellent fault tolerance mechanisms for component-level failures [60]. For computer scientists, there are tremendous similarities between neural systems and digital systems. Components such as cell body, dendrites, axons, nerve terminals, and synapses together constitute a neuron unit. Specifically, the core part of the neuron is a cell body containing a nucleus, with a radius of 2 to 60 microns; there are two types of cell processes of different lengths on the surface of the cell body, which are long axons (only one) and short dendrites (usually multiple); the excitatory transmission between neurons passes through axons and nerve terminals and finally reaches the synapse (the connection point between neurons) [61]. Neurons with various functions constitute a complete nervous system, which can effectively receive, integrate and transmit information/signal. This is considered to be the core link in the process of nervous system learning and adaptation. Although brain neural networks have different information processing and logical analysis capabilities at different levels, they are a coordinated and unified whole and are closely connected with each other [61]. The neuromorphic computer is a novel computer model that simulates the operation of the brain's neural network with ultra large-scale pulses and real-time communication [62]. Neuromorphic computers simulate the high performance, low power consumption, real-time and other characteristics of biological brain neural networks, and use large-scale CPU/GPU clusters to implement neural networks. In the CPU cluster, each thread will map and simulate the corresponding neuron, and thousands of threads (neurons) run in an orderly coordinated manner to form a complete large-scale neural network.

The implementation of neuromorphic computing at the hardware level can be achieved through oxide-based (CMOS) devices such as memristors, spin electronic memories, threshold switches, transistors, etc. Some examples are shown in Figure 8. Back in 2006, researchers at Georgia Tech proposed a field-programmable neural array [57]. The chip is the first in a series of increasingly complex floating-gate transistor arrays that allow the charge on the MOSFET gate to be programmed to simulate the channel ion properties of neurons in the brain, and is the first silicon programmable Neuron array. At the same time, many transistor-based brain-inspired chips and brain-inspired computer systems have also developed to a certain extent. For example, Stanford University developed the brain-inspired chip "Neurogrid" based on analog circuits, the University of Manchester began to develop SpiNNaker, a multi-core supercomputer that supports spiking neural networks based on ARM, the European Union's FACETS (fast analog computing with emergent transient states) project, and the U.S. Defense Research Agency DARPA's SyNAPSE (systems of neuromorphic adaptive plastic scalable electronics) project [55, 57, 63–65]. In 2008, HP realized a memristor prototype that

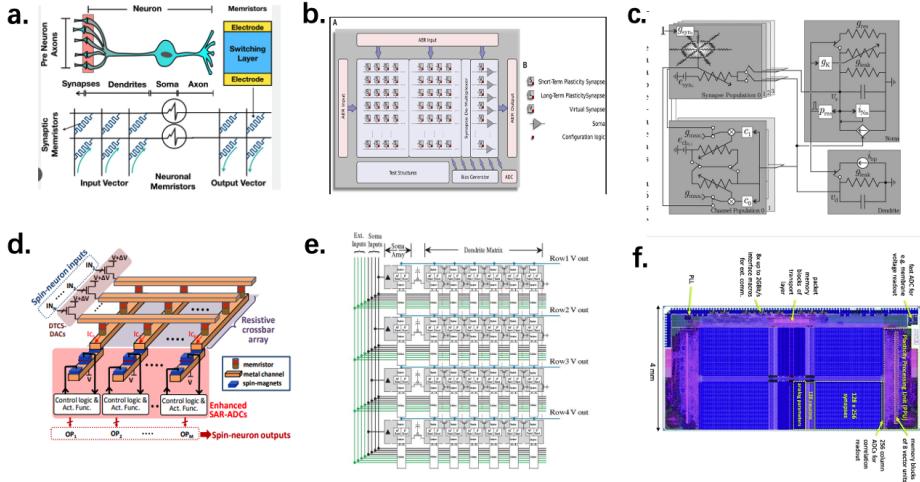


Figure 8: Conventional neuromorphic computing systems, utilizing electronics rather than photonics. a. Neuromorphic Computing with Memristor Crossbar [53]; b. A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128K synapses [54]. Block diagram of the architecture, showing two distinct synapse arrays (short-term plasticity and long-term plasticity synapses), an additional row of synapses (virtual synapses) and a row of neurons (somas); c. Neurogrid: A mixed-analog-digital multi-chip system for large-scale neural simulations [55]; d. SPINdle: SPINtronic deep learning engine for large-scale neuromorphic computing [56]. Shows an array of M spin-neurons that take N inputs each. The spin-neuron array consists of (i) Deep Triode Current Source Digital to Analog Converters that convert the N digital inputs into analog currents, (ii) a resistive crossbar array ($N \times M$) that is used to perform weighted summation of the neuron inputs, and (iii) enhanced Successive Approximation Register ADCs that evaluate the activation function and produce the M digital outputs; e. A field programmable neural array (FPNA) [57]. Each node of the matrix is identical, with the exception of the triangle wave generator on the output of the soma. Each node has some circuitry for readout, 1 Na^+ channel, 1 K^+ channel, 1 inhibitory synapse, and 1 excitatory synapse. Each node of the dendrite matrix is connected to its nearest neighbor in 2 dimensions; f. Layout of the current BrainScaleS-2 full-size ASIC [58]. It contains 512 neuron circuits and 131072 synapse circuits which are arranged in 4 quadrants. Data lines of the synapse arrays and the column ADCs are directly connected to the PPUs at the top and bottom edges. Each PPU contains 8 vector units with dedicated memory blocks in addition to the general-purpose processor part. Panels reproduced with permission from: a. Reproduced with permission.[53] Copyright 2018, Wiley. b. Reproduced under the terms of the Creative Commons Attribution License (CC BY) (<http://creativecommons.org/licenses/by/4.0/>).[54] Copyright 2015, the Authors. c. Reproduced with permission.[55] Copyright 2014, IEEE. d. Reproduced with permission.[56] Copyright 2014, ACM. e. Reproduced with permission.[57] Copyright 2006, IEEE. f. Reproduced under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>).[54] Copyright 2020, the Authors and Springer.

could simulate the function of synapses, and demonstrated the first hybrid circuit of memristor and silicon materials [66]. The global craze for artificial synapses began to take off in June 2012, when spintronics researchers at Purdue University published a paper on designing neuromorphic chips using lateral spin valves and memristors [56]. They believe the architecture works similarly to neurons and could therefore be used to test methods of reproducing brain processing. In addition, these chips are significantly more energy-efficient than conventional chips. In the same year Dr. Thomas and his colleagues at the University of Bielefeld created a memristor with learning capabilities. And in the following years [67], Thomas used this memristor as a key component of an artificial brain. Because of this similarity between memristors and synapses, it is an excellent material for building artificial brains—and thus a new generation of computers. Memristor allows us to build extremely energy-efficient, durable, and self-learning processors. It is precisely because of this ability that memristors have also received increasing attention in areas like neural networks and artificial

intelligence. Several research groups are now exploring the use of memristors to build more efficient neural network architectures.

2.3 Challenges and limitations faced by conventional neuromorphic computing

The limitations of traditional brain-inspired computing mainly encompass three aspects: hardware limitations, software and algorithm limitations, and practicality and application limitations. They will be discussed in detail below in the text as well as in Figure 9, which compares biological brains, electronic, and photonic neuromorphic computing systems.

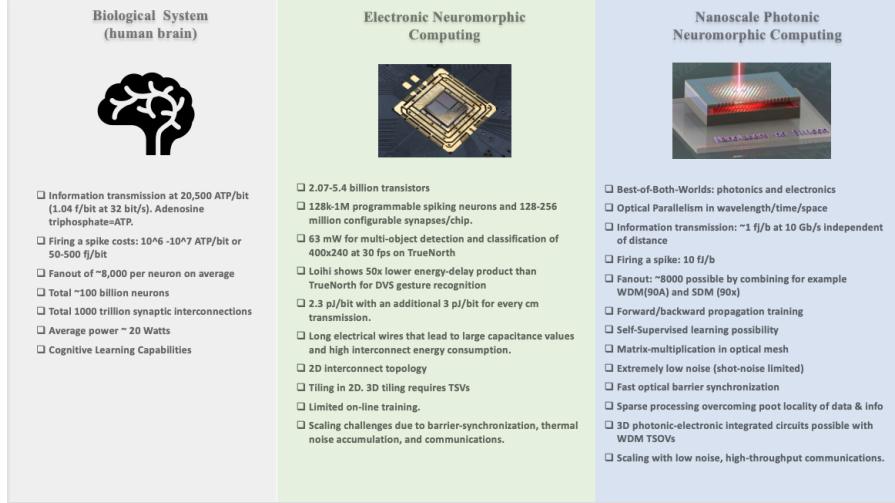


Figure 9: Comparisons of a biological cognitive system (human's cortex and brain, Figure 2a), CMOS-based electronic neuromorphic computing (E.g., IBM TrueNorth, Intel Loihi, Figure 8), and the latest photonic neuromorphic computing (Figure 3 and Table 1). Data in this figure are extracted from: [68–72]. Reproduced under the terms of the Creative Commons Attribution License (CC BY) (<http://creativecommons.org/licenses/by/4.0/>).[73] Copyright 2022, the Authors and AIP Publishing.

Although neuromorphic computing advertises low power consumption, existing hardware typically consumes more energy than biological nervous systems. Decades of research and billions of dollars have been invested in various forms of pattern recognition, and while substantial improvements have been made, synthetic electronic systems still fall short of the capabilities of human perception on specific problems [74, 75]. Materials play a key role in the energy consumption of neuromorphic computing. Due to the conductivity and resistivity characteristics of conventional conductors, a large amount of energy loss will inevitably occur during the transmission of electrical currents. Moreover, the size factor is also limiting the energy consumption limit of traditional conductive materials. As neuromorphic computing devices continue to shrink in size, quantum and thermal effects may become more pronounced, affecting energy efficiency [5, 76]. As seen in Figure 9, conventional electronic systems typically consume three orders of magnitude more energy than biological systems and photonic systems.

In computing systems and networks, "Communication Overhead" is a term used to describe the additional resources and time required for information transmission [77]. It is an important factor affecting the computing performance and energy efficiency. For neuromorphic computing, this concept has several important aspects. First, data transmission is a core issue because simulating neurons in neural networks requires information transfer through synapses, which usually consumes additional energy and time in hardware implementation [78]. Second, the activity of neurons usually needs to be represented through specific encoding and decoding methods, which also increases computational and energy overhead. In addition, some neuromorphic computing models require precise timing synchronization between neurons, which further increases communication and computing overhead [79, 80]. In large-scale neural networks, information also requires complex routing between multiple neurons and network layers, which often requires additional hardware and algorithm support. In order to ensure the accuracy of information transmission, error detection and correction mechanisms need to be introduced, which will also bring more communication overhead [81]. Finally, under high load

or high concurrency, information may need to be queued in a buffer waiting to be processed, resulting in latency and additional energy consumption [82, 83]. Therefore, communication overhead is an important factor affecting the performance and energy efficiency of brain-inspired computing, and its optimization usually requires comprehensive consideration from multiple aspects such as hardware design, network topology, and signal encoding methods. In Figure 9, it's shown that photonic systems can transmit data at a high rate and a low energy level independent of distance, which essentially solves the communication overhead of electronic systems.

Parallelism and synchronization also influence power consumption of neuromorphic computing. In traditional electronics, parallelism and synchronization face several major limitations. First, in terms of parallelism, electronic devices are often limited by circuit bandwidth, which affects their ability to perform highly parallel processing [84, 85]. Additionally, in highly parallel electronic systems, electromagnetic interference can become a serious problem, limiting overall performance [72]. Second, in terms of synchronization, maintaining global clock synchronization can be very complex and energy-intensive in large-scale electronic systems [86]. In addition, the propagation speed of electronic signals in wires is limited, which further affects the accuracy of system synchronization. Therefore, these factors together constitute the main limitations of traditional electronic devices in terms of parallelism and synchronization. Photonic systems, on the other hand, can achieve high-degree of parallelism using the inherent fanout and WDM techniques , depicted in Figure 9.

In addition to the limitations of energy consumption of conventional neuromorphic computing, imitating large-scale neural networks requires a large amount of hardware resources, which is impractical in terms of space and cost.

To mimic large-scale biological neural networks, large numbers of neurons and synapses are needed. This results in larger hardware size or footprint, which increases cost and space requirements. For example, BrainScaleS is a neuromorphic computing system for large-scale simulations, which requires a full-sized dedicated computer room to accommodate it [58]. Neurogrid, A hardware platform that simulates cortical neural networks, comes in a size equivalent to a standard 19-inch rack [55]. On the other hand, quantum effects make it impossible to further reduce the size of transistors and therefore the size of brain-like computers [52, 87]. Moreover, the complex interconnection structure between neurons faces huge challenges in hardware implementation, especially when pursuing low power consumption and high performance. Specifically, neuromorphic computing involves a series of complex dynamics and nonlinear operations, which undoubtedly increases the complexity of hardware and algorithm design. At the same time, simulating biological neural networks usually requires storing a large amount of parameters and status information, which not only places higher requirements on storage resources, but also has a considerable impact on system size and energy consumption [58]. As the network scale expands, how to effectively expand hardware scaling and maintain low energy consumption becomes particularly critical. In this context, integrating multiple components such as neurons, synapses, and learning rules into an efficient operating system becomes a difficult task. Finally, it is worth noting that existing semiconductor processing technology also has certain limitations in terms of integration density, power consumption and reliability [88]. Therefore, solving these problems of size and complexity requires interdisciplinary research collaboration and continued technological innovation in the industry.

Last but not least, in neuromorphic computing, software and algorithms face a series of challenges and limitations as well. First, the software implementing these models is often extremely complex and computationally intensive due to the complex neurodynamics and nonlinear calculations involved [6]. This complexity not only limits the feasibility of the algorithm in practical applications, but also increases the computational burden. Secondly, especially in applications based on spiking neural networks (SNNs), the algorithm training process often requires a large amount of time and computing resources, which poses an obvious obstacle to scenarios that require real-time or near-real-time response [79]. Furthermore, compared with traditional machine learning algorithms, neuromorphic algorithms are often more difficult to interpret and verify, which is particularly problematic in applications that are highly sensitive and require interpretability, such as medical or autonomous driving [89]. In addition, most of the existing brain-inspired algorithms are developed on traditional computing platforms that do not fully match the brain-inspired hardware architecture. This mismatch often leads to algorithm performance degradation in real hardware environments. At the same time, although the learning rules and plasticity mechanisms in biological neural networks are extremely complex, current algorithms fail to fully simulate these advanced characteristics, thus limiting their effectiveness in handling more complex real-world tasks [62]. Finally, although brain-like algorithms

perform well on certain specific tasks, they usually lack the versatility and adaptability to demonstrate across multiple tasks and different environments.

3 Overview of the Fundamentals of Photonics

3.1 Basics of photonics and light-matter interactions

Photonics is a multidisciplinary field that studies light propagation and light–matter interactions, which has applications in various fields such as optical sensing [90], optical interconnection, and optical communication [91]. Photonics also plays an important role in developing optical computing systems, providing an alternative to electronic computers for high-speed and low-power data processing and transmission [92, 93]. This section introduces the basics of photonics in light-matter interactions.

Maxwell's equations constitute the cornerstone of classical electrodynamics, delineating the fundamental principles underlying the dynamical behavior of electric and magnetic fields. These equations encapsulate the generation of electromagnetic fields by charges and currents (Gauss's law for electricity and Ampère's law with Maxwell's addition), the non-existence of magnetic monopoles (Gauss's law for magnetism), and the induction of electric fields by time-varying magnetic fields (Faraday's law of induction). In the context of light-matter interactions, Maxwell's equations are indispensable. They predict the self-propagating nature of electromagnetic waves in a vacuum, a phenomenon in which light is a principal exemplar. Upon interaction with matter, these equations govern the complex processes of reflection, refraction, absorption, and transmission. The boundary conditions dictated by Maxwell's equations at interfaces between different media determine the electromagnetic field distributions and, consequently, the optical responses of materials. Furthermore, Maxwell's equations are integral to the quantification of the dielectric and magnetic properties of materials, characterized by permittivity and permeability, respectively. These properties influence the phase velocity of electromagnetic waves in media, leading to phenomena such as dispersion and polarization. The equations also describe the conservation of energy and momentum in electromagnetic fields, providing a framework for understanding the interaction forces and torques exerted by light on material particles, which is the basis for optical trapping and manipulation technologies.

Equations (i)–(iv) are Maxwell's equations that dictate the behavior electromagnetic waves:

$$\begin{array}{ll} \nabla \cdot \mathbf{D} = 4\pi\rho_f & \text{(i)} \\ \nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} & \text{(iii)} \\ \nabla \cdot \mathbf{B} = 0 & \text{(ii)} \\ \nabla \times \mathbf{H} = \mathbf{J}_f + \frac{\partial \mathbf{D}}{\partial t} & \text{(iv)} \end{array}$$

where \mathbf{D} is electric flux density, \mathbf{B} is magnetic flux density, \mathbf{E} is electric field intensity, and \mathbf{H} is magnetic field intensity. Maxwell's equations form the foundation of classical photonics and light-matter interactions.

To build efficient optical neural networks (ONN), nanoscale components are required to manipulate and process light signals. Diffraction and interference of light at the nanoscale indeed play a crucial role in these components. These phenomena enable precise control over light signals, allowing for information encoding, processing, and transmission within the network. In principle, diffraction and interference both arise due to the superposition principle, which states that when two or more waves overlap, the resulting wave is the sum of their individual amplitudes. For example, the photonic oscillations E_1 and E_2 produced at a certain point by coherent light waves can be expressed as,

$$\begin{aligned} E_1 &= a_1 \exp [i(\alpha_1 - \omega t)] \\ E_2 &= a_2 \exp [i(\alpha_2 - \omega t)] \end{aligned}$$

So, it can be readily deduced that the resultant superimposed oscillation is

$$E_s = a \exp [i(\alpha - \omega t)]$$

Where

$$a^2 = a_1^2 + a_2^2 + 2a_1a_2 \cos(\alpha_2 - \alpha_1)$$

$$\tan\alpha = \frac{a_1 \sin\alpha_1 + a_2 \sin\alpha_2}{a_1 \cos\alpha_1 + a_2 \cos\alpha_2}$$

It is obvious that by varying α_1 and α_2 , both the amplitude a and phase α of the combined light wave are changed. By architecting specialized dielectric material structures on a nanoscale, intricate photonic interactions can be actualized.

For example, B. Wu et. al. actualized a variety of nonlinear activation functions through the employment of the thermo-optic effect and micro-ring resonators [94]. They use germanium whose absorption coefficient α and refractive index n are characterized by temperature T as follows:

$$\begin{aligned}\Delta\alpha &= k_1 \exp(\Delta T) \\ \Delta n &= k_2 \Delta T\end{aligned}$$

Where k_1 and k_2 are constant and ΔT is change of temperature. These two values have influence on quality factor of micro-ring resonators, and thus change the output power of it. Another example is MZI, which is a device that utilizes the principle of light interference to measure small changes in phase or refractive index [95]. It consists of a beam splitter that splits an incoming light beam into two paths, which then recombine at a second beam splitter. However, we can also utilize the electro-optic effect and nonlinear effects to alter the refractive index and phase, modify the interference conditions in MZI, and finally realize the change of light wave amplitude.

3.2 Comparison between photons and electrons

Electrons and photonics are both elementary constituent particles of the physical world. Electrons are fermions which are particles with half-integer spins and they obey Pauli's exclusion principle. This means that no two fermions can occupy the same energy state. Other fermions include: Protons, Neutrons, Neutrinos, Quarks. Fermions are different from bosons, which have integer spins and do not obey Pauli's exclusion principle. Some examples of bosons include: photons, a-particles, Higgs, Helium atoms, and Gluons. More quantitatively, bosons are the fundamental particles that have spin in integer values (0, 1, 2, etc.). Fermions, on the other hand, have spin in odd half integer values (1/2, 3/2, and 5/2, but not 2/2 or 6/2).

Wave function of electrons is prescribed by the Schrodinger equation (Equation v), while that of photons is derived from the Maxwell's equations in section 3.1 and is sometimes referred to as the plane wave solutions (Equation vi-vii). Electrons have a specific mass and a drift velocity of limited magnitude, whereas photons are massless and travel at the speed of light.

The Schrödinger equation for the electron is:

$$E\psi = -\frac{\hbar^2}{2\mu} \nabla^2 \psi - \frac{q^2}{4\pi\epsilon_0 r} \psi \quad (v)$$

where E is energy, q is the electron charge, \mathbf{r} is the position of the electron relative to the nucleus, $r = |\mathbf{r}|$ is the magnitude of the relative position, the potential term is due to the Coulomb interaction, wherein ϵ_0 is the permittivity of free space and

$$\mu = \frac{m_q m_p}{m_q + m_p}$$

is the 2-body reduced mass of the hydrogen nucleus (just a proton) of mass m_p and the electron of mass m_q .

The planar traveling wave solutions of the photon wave equations are:

$$\mathbf{E}(\mathbf{r}) = \mathbf{E}_0 e^{-i\mathbf{k}\cdot\mathbf{r}} \quad (vi)$$

$$\mathbf{B}(\mathbf{r}) = \mathbf{B}_0 e^{-i\mathbf{k}\cdot\mathbf{r}} \quad (vii)$$

where $\mathbf{r} = (x, y, z)$ is the position vector (in meters) and \mathbf{k} is the wavenumber.

In semiconductor physics, electrons and photons are closely intertwined. A diode laser or a light-emitting diode (LED) is a semiconductor device that emits light when electrical current flows through

it. As A. Einstein correctly predicted in 1916, electrons in the semiconductor can recombine with electron holes, releasing energy in the form of photons. The color of the emitted light (corresponding to the energy of the photons) is determined by the energy required for electrons to cross the band gap of the semiconductor.

4 Photonic Components for Neuromorphic Computing

4.1 Key devices of photonics for neuromorphic computing

Nano-photonics, as an emerging interdisciplinary subject, integrates the principles of nanotechnology and photonics, aiming to explore and harness the manipulation of light wave by nanoscale structures. In the landscape of photonics, active devices and passive devices are crucial and have broad application prospects. Neuromorphic systems aim to mimic the computational and cognitive capabilities of the brain by leveraging the principles of neural networks. In this part, we will focus on the key devices in photonics, divided into two parts: active devices and passive devices, and provide an in-depth analysis of their working principles, applications in neuromorphic computing, and future development trends. Some representative devices are shown in Figure 10.

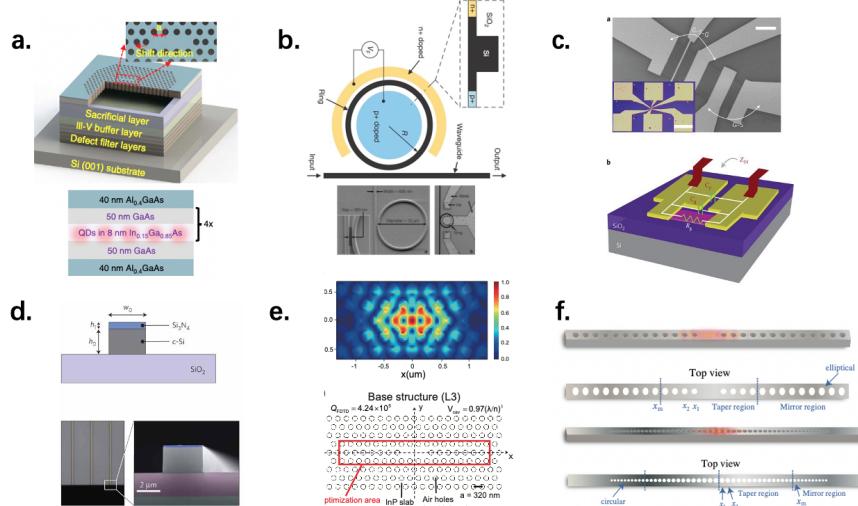


Figure 10: Key photonic devices and structures for neuromorphic computing. a. Schematic diagram of the fabricated InAs/GaAs QD PC (L3 cavity) laser epitaxially grown on on-axis Si (001) substrate [96]. The lattice constant, radius and shift of L3 PC cavity are a , r and $0.15a$, respectively. b. Schematic layout of the ring resonator-based modulator [97]. The inset shows the cross-section of the ring. R , radius of ring. VF , voltage applied on the modulator. SEM and microscope images of the fabricated device. c. SEM and optical (inset) images of the high-bandwidth graphene photodetectors [98]. The graphene shown here has two to three layers. Two types of wirings are shown: ground–signal (G–S) and ground–signal–ground (G–S–G). The high-frequency results are from devices with G–S wirings. d. Schematic representation of the cross-section of a strained SOI waveguide [99]. Si forms the core of the waveguide and the parameters are chosen to have more than 95% of the optical field confined within the waveguide. e. The L3 photonic crystal resonance nanocavity [100]. f. The nanobeam photonic crystal structure [101]. Panels reproduced with permission from: a. Reproduced under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>).[96] Copyright 2020, the Authors and Springer Nature. b. Reproduced with permission.[97] Copyright 2005, Springer Nature. c. Reproduced with permission.[98] Copyright 2009, Springer Nature. d. Reproduced with permission.[99] Copyright 2012, Springer Nature. e. Reproduced under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>).[100] Copyright 2021, the Authors and Optica. f. Reproduced under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>).[101] Copyright 2023, the Authors and De Gruyter.

1. Active devices

Active devices are photonic devices that can generate, manipulate or amplify optical signals, including lasers, optical modulators, photodetectors, etc. [102] These devices play a vital role in fields such as communications, sensing, laser processing, and medical diagnosis.

1.1 Laser

Lasers are the most important and widely used active devices in photonics and are known for their highly coherent light output, high emission power, and high spectral brightness [96]. In the current state of photonics, different types of lasers such as Vertical Cavity Surface Emitting Lasers (VCSELs), DFB lasers, micro-ring lasers, quantum dot lasers, and the latest topological lasers and PCSELs (see section 6.1 for details) [103, 104] are developing rapidly [105]. These lasers not only play a key role in optical communications, but are also used in fields such as Lidar, optical sensing and medical imaging. Lasers can be pulsed, continuous-wave, optically or electrically pumped, depending on their specific needs.

1.2 Optical modulator

Light modulators are used to control the intensity, phase and frequency of light in real time [97]. In photonics, technologies such as electro-optic modulators and acousto-optic modulators are widely used in high-speed optical communication systems to achieve high-speed data transmission and information processing. Nanoscale optical modulators offer the advantages of small size, low power consumption and high speed and are key components of next-generation communication systems and photonic computers [106].

1.3 Light detector

Photodetectors are important active devices of photonics, used to convert optical signals into electrical signals [98]. There is a wide variety of photodetectors available, including photodiodes, photodetection arrays, and single-photon detectors. These devices play pivotal roles in communications, sensing, optical imaging, and quantum information processing for next-gen photonic computing systems [107, 108].

1.4 Future Prospects of Active Devices

In the future, active devices will continue to grow and expand in photonics. The miniaturization, low power consumption and high efficiency of lasers will empower more applications, including quantum information processing and PICs. Optical modulators will continue to play a critical role in high-speed communications and Lidar. Photodetectors will usher in higher sensitivity, faster response times, and wider wavelength ranges, driving innovation in fields such as wireless communications, medical diagnosis, and optical sensing. All these devices will substantially contribute to the growth of neuromorphic computers in the near future.

2. Passive components

Passive devices refer to photonic devices that cannot generate or amplify optical signals, but they are equally essential for the transmission, control and processing of light. Common passive devices include waveguides, resonators, photonic crystals, etc.

2.1 Waveguide

A waveguide is a structure used to guide light in a desired direction, usually consisting of a high-refractive index material surrounded by a low-refractive index cladding [99]. There are many types of waveguides, including optical fiber waveguides, planar waveguides, photonic crystal waveguides, etc. They play a key role in optical communications, optical sensing, photonic chips and other fields [109]. The low transmission loss of fiber optic waveguides makes them ideal for long-distance communications, while planar waveguides are often used in optical chips to realize micro-nanoscale integrated optical components [110].

2.2 Resonant cavity

A resonant cavity is a device used to enhance light modes of specific frequencies [100]. There are many types of resonant cavities, including fiber cavities, micro-ring cavities, micro-beam cavities, etc. Resonators can be used in filtering, lasing and sensing applications, and their high Q-factors give them a special place in photonics. Micro-ring cavities achieve resonance through multiple

photon reflections and are widely used in micro-lasers, sensors and quantum information processing [111–113]. They can also be used along with soliton microcombs to perform wavelength division multiplexing. A nanobeam cavity [114] is an elongated waveguide with a micron-scale cross-section that can be used to achieve a high-quality factor resonator whose size and shape can be tuned for specific applications. Key performance metrics of resonant cavities are the Q-factor and modal volume.

2.3 Photonic crystal

Photonic crystal is a material with a periodic structure that can realize the light band gap and waveguide mode of light [115, 116]. It has a periodically varying refractive index enabled by an array of holes or rods. These band gap and waveguide modes can be used in optical filtering, lasers, sensors, and other applications [117, 118]. Photonic crystal design and preparation technology plays an important role in laser designs, such as the popular Photonic Crystal Surface Emitting Lasers (PCSEL) [104] and the more recent topological insulator lasers.

2.4 Future prospects of passive components

In the future, the development of passive devices will continue to promote the development of neuromorphic computing technologies. The design of the waveguide will be more flexible to meet the needs of different applications. The resonant cavity will further improve the quality factor and allow optical devices to operate more efficiently. The structure and function of photonic crystals will be further expanded to meet the requirements of a wider range of engineering applications. The continued development of these passive devices will bring more possibilities to the future of photonics-enabled areas.

We can agree that active devices such as lasers, optical modulators, and photodetectors provide optical signal generation and processing capabilities, supporting applications such as optical communications, medical imaging, and quantum information processing. Passive devices such as waveguides, resonators and photonic crystals provide key tools for light transmission and manipulation, supporting the development of optical communications, sensing applications and data transmission. In the future, these key devices will continue to be enhanced while photonics continues to drive cutting-edge research in neuromorphic computing, opening up new possibilities for future scientific and engineering applications.

4.2 Optical interconnects for scalable neuromorphic systems

Neuromorphic systems, inspired by the structure and function of the brain, have attracted significant attention in the AI and computational neuroscience community. Neuromorphic systems aim to mimic the computational and cognitive capabilities of the brain by leveraging the principles of neural networks. As the complexity and scale of these systems increase, the need for efficient, high-bandwidth connectivity becomes increasingly important, because providing efficient and scalable connections to support the massive flow of data between neurons and synapses is critical [119]. Traditional electronic interconnects, despite their widespread use, still face bottlenecks in bandwidth, power consumption, crosstalk, and latency as neuromorphic systems evolve. Optical interconnects, which use photons to transmit information, have emerged as a promising solution to address the limitations of traditional electronic interconnects in neuromorphic computing. In this part, we provide an overview of recent advances in optical interconnects for scalable neural systems, highlighting key advances, challenges, and prospects. We draw insights from numerous research papers, focusing on advances in photonic components, network topologies, and system integration. Some representative optical interconnects are shown in Figure 11.

4.2.1 Advancements in Photonic Components

The photonic components are fundamental components of optical interconnects. Recent innovations in this field have significantly improved the performance and efficiency of optical interconnects for neuromorphic systems.

Photonic integrated circuits (PIC)

Photonic integrated circuits are compact and complex chips that combine multiple optical functions on a single chip. From the Shannon-Hartley theorem, PICs have an unlimited bandwidth $\approx 40\text{THz}$ which indicates a promising future of PICs to transport and process analogue signals. Recent developments

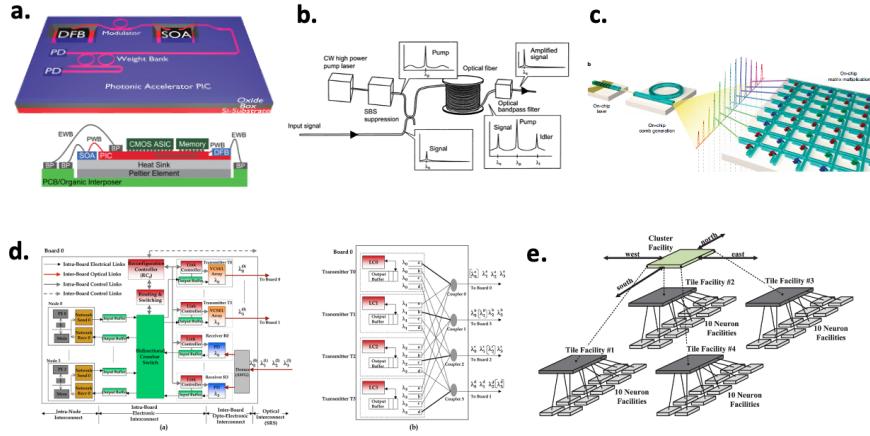


Figure 11: Optical interconnects for scalable neuromorphic systems. a. Neuromorphic photonic accelerator PIC [120]. Prospect of a packaging solution consisting of a silicon PIC, CMOS chip, and III-V DFB lasers and SOAs integrated via photonic wire bonds and multi-chip integration. b. General scheme of phase-insensitive fiber-based optical parametric amplifier [121]. c. Photonic tensor core [15]. Conceptual illustration of a fully integrated photonic architecture to compute convolutional operations. An on-chip laser (not included here) pumps an integrated Si_3N_4 microresonator to generate a broadband soliton frequency comb. d. Proposed on-board interconnect for the E-RAPID architecture with RC and LCs and proposed technology for reconfiguration using passive couplers and array of lasers per transmitter port [122]. e. Proposed scalable NoC architecture for implementing clusters of neurons to implement spiking neural networks [123]. It aims to address the scalability issue by creating a modular array of clusters of neurons using a hierarchical structure of low and high-level routers. Panels reproduced with permission from: a. Reproduced with permission.[120] Copyright 2022, IEEE. b. Reproduced with permission.[121] Copyright 2002, IEEE. c. Reproduced with permission.[15] Copyright 2021, Springer Nature. d. Reproduced with permission.[122] Copyright 2010, IEEE. e. Reproduced with permission.[123] Copyright 2012, IEEE.

in PICs have led to improved signal processing capabilities, allowing the implementation of complex neural network architectures. PICs also enable efficient signal routing, switching, and conditioning in neuromorphic systems [120, 124].

Nonlinear Optical Devices

Nonlinear optical devices, such as optical parametric amplifiers [121] and inverters, have found application in improving the efficiency of optical interconnects. They enable wavelength conversion and amplification of optical signals, reducing the need for power-consuming electronic repeaters [125].

Photonic Neural Network Accelerators

Research is underway to develop specialized photonic devices to accelerate neural network (deep learning) computations. Optical processors based on metasurface layers, soliton microcombs, VC-SELs, and ring resonators can perform MVM at extremely high speeds, significantly improving the efficiency of deep learning in neuromorphic systems [7, 15, 78].

4.2.2 Evolution of network topology

Network topology in neuromorphic systems is essential for efficient learning and information transfer. Recent research has explored new optical network architectures to improve scalability and system performance.

Reconfigurable optical interconnects

Reconfigurable optical interconnects provide flexibility in neural network configuration. By dynamically changing the connections between neurons and synapses, these connections enable rapid adaptation and learning [122, 126–128]. Research has focused on developing tunable optical components and switches to facilitate reconfiguration.

Neural Network-on-Chip (NoC)

Neural NoCs are specialized optical interconnects designed to efficiently connect various processing elements, including neurons and synapses. They reduce communication congestion and provide low-latency, high-bandwidth connections [123]. Recent developments have demonstrated the benefits of NoC in large-scale nervous systems.

4.2.3 System Integration and Challenges

The integration of optical interconnects into neuromorphic systems poses several challenges related to compatibility, scalability, and power consumption.

Hybrid Integration with Electronics

Integrating optical interconnects with existing electronic neuromorphic systems is a complex task. Researchers are working on hybrid integration techniques to ensure seamless communication between electronic and photonic components, maximizing the benefits of both technologies [129, 130]. Heterogeneous and monolithic integration of semiconductor materials in optical interconnects is a branch of technique that is alternative to hybrid integration.

Scalability

Scalability is a key challenge for optical interconnects. As neuromorphic systems grow in size, ensuring that optical components can be manufactured and interconnected at scale without a significant increase in cost and complexity is essential [131, 132].

Power consumption

Although optical connections are more energy efficient than electronic connections, minimizing power consumption is still a top priority. Research is ongoing to develop low-power photonic components and advanced power management techniques [86] for neuromorphic systems.

In short, optical interconnects have the potential to revolutionize neuromorphic computing by providing scalable, high-bandwidth, and energy-efficient communications paths. As research on photonic components, network topologies, and system integration continues to bloom, we can expect significant advances in the development of large-scale neuromorphic systems. Integrating optics into neuromorphic hardware promises to accelerate advances in AI, cognitive computing and a range of applications, including robotics, autonomous vehicles and biomedical research. Recent advances in these three areas discussed above are paving the way for efficient and scalable neuromorphic systems. While challenges regarding compatibility, scalability, and power consumption remain, the promise of optical interconnects to revolutionize neuromorphic computing is becoming increasingly clear.

4.3 Optical logic gates in neuromorphic computing

Similar to electronic computing, in optical computing, logic gate serves as the very bedrock upon which systems are constructed. In recent years, notable advancements have been achieved for optical logic gates, which have paved the way for enhanced neuromorphic computing [133]. The realization of optical logic gates have many different strategies, such as diffractive neural networks [13], semiconductor optical amplifiers [133–135] and photonic crystal waveguide [101].

Diffraction neural networks largely rely on the physical phenomenon of diffraction for computation. Typically, these networks undergo a pre-training process wherein they learn the intricate relationship between the input light field and the resulting patterns. This learning procedure is accomplished by adjusting the phase delay or transmission rate at each specific point in the network. Diffraction neural networks are generally composed of multiple metasurfaces [13, 136, 137], each layer equating to a layer of neurons. With suitable training, manufacturing the corresponding diffractive network layers, and assembling them rightly, it can execute the required operations on any given input light field in real-time while performing neural network calculations at the speed of light. Diffraction neural networks operating in the mid-infrared are anticipated to function within a scale of a few millimeters. This endows it with immense prospects for large-scale integrated applications.

Semiconductor optical amplifiers (SOAs) are used to form optical logic gates due to their ability to modulate the intensity of light, change the phase of light, or both [134]. They can perform several different logical operations based on the principle of gain saturation, where the output light intensity is dependent on the input light intensity. This characteristic makes SOA highly valuable for

building optical logic gates, enabling it to implement various applications such as all optical lattices, pseudo-random bit sequence generation, and all optical encryption.

Photonic crystal waveguides are structures that can limit and guide light in periodic dielectric media. They can control the propagation of light in a highly precise manner, making them an excellent platform for implementing optical logic gates. Especially in recent years, the emerging topological photonic crystal waveguides [138–141] have significant advantages in the construction of all-optical logic gates due to their unidirectional propagation, controllable splitting, and other attractive characteristics.

These advancements in optical logic gates above have made optical neuromorphic computing a promising field for the development of high-speed, low-energy, and compact computing systems. In more recent years, researchers have begun to explore more sophisticated photonic devices such as frequency microcombs [15], micro-ring resonators [16], phase change materials [15], nanolasers [17], metasurfaces [18], optical attenuators [19], EDFA [16], and photodiodes [19]] to achieve state-of-the-art neuromorphic computing systems. These will be discussed in detail in the remaining text.

4.4 State-of-the-art fabrication platforms for photonic devices

Nano-photonics is a multidisciplinary field concerned with manipulating light at the nanoscale, often on the order of the wavelength of light. This field has grown rapidly due to its potential applications in telecommunications, sensing, imaging, and computing. Photonic devices offer advantages such as higher speed, lower power consumption, and compact size, but making these devices requires advanced manufacturing techniques capable of the precise drilling and etching of fine nanostructures. In this part, we provide an overview of the state-of-the-art manufacturing platforms for photonic devices. We explore recent developments in lithography, etching, self-assembly and 3D printing, demonstrating their potential to revolutionize photonic technologies.

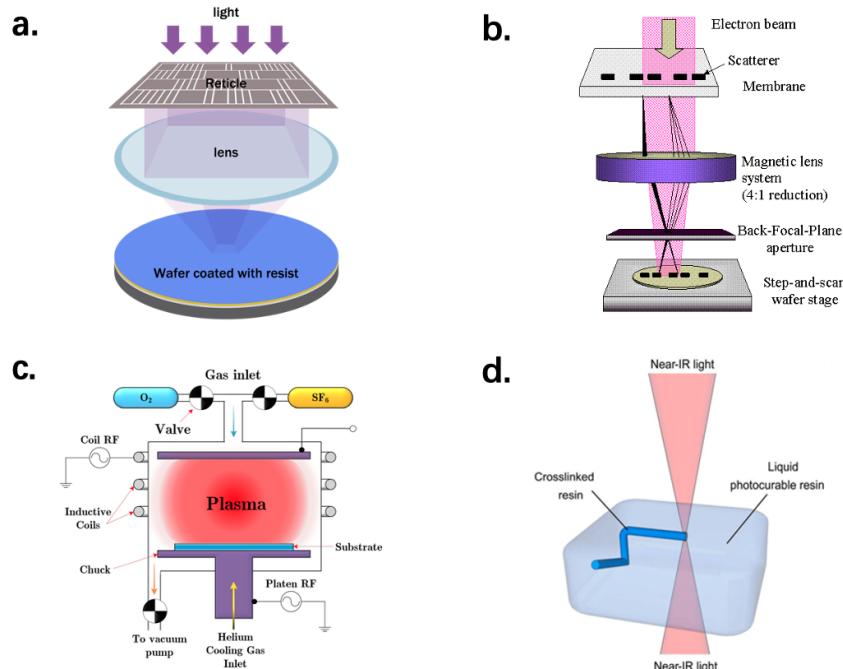


Figure 12: State-of-the-art fabrication methods for photonic devices. a. Photolithography. b. Electron beam lithography (EBL). c. ICP-RIE etching. d. 3D printing with two-photon polymerization (TPP). Panels reproduced with permission from: c. Reproduced with permission.[142] Copyright 2017, AIP Publishing. d. Reproduced under the terms of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).[143] Copyright 2020, the Authors and MDPI.

4.4.1 Lithography

Lithography plays an important role in the production of optical devices and includes a variety of methods, including traditional photolithography, extreme ultraviolet (EUV) lithography, electron beam exposure and nano-printing technology.

First of all, traditional photolithography [144] is a popular technology for manufacturing optical devices. It uses a UV light source and a mask to expose the photoresist, then transfers the pattern to the optical material by chemical etching. This technology is widely used to create microscopic arrays, waveguides, sensors and other optical components. With the continuous improvement of equipment, traditional photolithography techniques have achieved higher resolution and alignment accuracy [145]. EUV lithography is an innovative method that uses an ultraviolet light source for exposure. EUV lithography has extremely high resolution and can handle more complex structures. It is widely used in the semiconductor industry but also has great potential in the production of optical devices, such as micro-lasers and optical communication components.

Electron beam exposure (EBL) technology is a high-resolution preparation method that uses a precise electron beam to expose the photoresist. This method is often used to make optical devices that require extremely high precision and fine patterns, such as terahertz detectors and quantum dot arrays. EBL technology can achieve a precision of less than 100 nanometers, which is very beneficial for the fabrication of fine structures [146].

In addition, nanoimprinting is an emerging method that reproduces micro- and nanostructures by applying pressure to the material surface. This technique is suitable for fabricating large-area structures such as nanoarrays and surface plasmonic resonators [147, 148]. This shows the potential of high-throughput manufacturing to increase production efficiency.

In short, the diversity of lithography techniques offers a multitude of tools and methods for fabricating optical devices. Whether traditional lithography, EUV lithography, electron beam exposure or nanoimprinting, they play a key role in innovations in photonics and optoelectronics, driving optical equipment with high performance, high resolution and high reliability. As state-of-the-art lithography techniques and tooling continue to improve, it is natural that photonic devices will become more precise and complex, supporting further development of optical communications, neuromorphic computing and laser applications.

4.4.2 Etching

Etching technology [149] is a key micro-nano processing method that is widely used for photonics and optoelectronics to accurately shape and regulate the surface and structure of optical materials. Reactive Ion Etching (RIE), for example, uses a combination of chemical and physical reactions to remove material from a substrate and is the simplest process that is capable of directional etching. A highly anisotropic etching process can be achieved in RIE through the application of energetic ion bombardment of the substrate during the plasma chemical etch. The RIE process thus provides the benefits of highly anisotropic etching due to the directionality of the ions bombarding the substrate surface as they get accelerated towards the negatively biased substrate, combined with high etch rates due to the chemical activity of the reactive species concurrently impinging on the substrate surfaces.

The main application areas of etching include waveguides, gratings, lasers, optical modulators, etc. First, etching technology plays a crucial role in waveguide fabrication. Waveguide is a basic component of optical signal transmission. Complex waveguide structures can be produced through etching technology, including straight waveguides, curved waveguides and waveguide gratings, thereby achieving efficient guidance and coupling of optical signals [150]. The preparation of gratings and nanolasers is also inseparable from the etching technology. Gratings are used for spectroscopy and spectral analysis. By making micron-scale periodic structures (like holes or rods) on semiconductor materials, optical signals of specific wavelengths can be separated [151]. Lastly, Lasers require high-precision preparation and delicacy, and etching can be used to precisely define the size and shape of the laser resonance cavity to ensure stable and efficient laser performance [152].

At the same time, optical modulators are indispensable devices in optical communications and optical signal processing, and are often prepared using etching technology. Etching techniques can be used to create gratings or modulation regions. By modulating the properties of these structures, modulation and control of optical signals, including amplitude, phase and frequency, can be achieved. In short, the application of etching technology in optical devices provides key support for the development of

photonics and optoelectronics. By precisely controlling optical structures, etching technology enables optical devices to achieve higher performance, greater scalability, and greater reliability, thereby driving advances in neuromorphic computing and photonic integrated circuits.

4.4.3 3D printing technology

3D printing technology [153], also known as additive manufacturing, has emerged in the production of optical devices, bringing many potential applications and innovations to the fields of photonics and optoelectronics. This technology uses a layer-by-layer method of stacking materials, allowing the fabrication of complex optical structures while providing a high degree of customization and personalization.

3D printing technology has many applications in the production of optical components. Complex optical components such as optical lenses, gratings, waveguides and reflectors can be precisely manufactured using 3D printing technology. This approach not only provides more flexible preparation but also reduces the number of optical components in the optical path, thereby reducing system complexity [154], and 3D printing technology also plays a key role in manufacturing micro-optical devices. Micro-optical components such as micro-lenses and micro-lens arrays can be manufactured with high precision using 3D printing technology to meet the needs of micro-optical systems [155, 156]. This is essential for applications such as micro cameras, biosensors and pico-projectors. In addition, optical waveguide fabrication also benefits from 3D printing, which can be used to create complex waveguide structures for efficient optical signal transmission. This is important for optical switches in data centers and optical communications because they deliver better performance and greater scalability. Therefore, by precisely controlling the arrangement and structure of materials, personalized photonic materials with special optical properties can be obtained [157, 158] for advanced computing tasks.

In short, 3D printing technology offers a new approach to manufacturing optical devices, with a high degree of flexibility and customization in the preparation process. This technology has made significant progress in the fabrication of optical components, micro-optical devices, optical waveguides and photonic materials, providing more opportunities for the development of photonics and optoelectronics. As 3D printing technology continues to develop and innovate, we will see its broader adoption in neuromorphic engineering.

4.4.4 Challenges and Future Directions

While remarkable progress has been made in photonic fabrication, several challenges remain:

Integration: efficient integration of various photonic components into functional systems, such as a PIC, is a complex task that requires interdisciplinary collaboration. Some existing integration methods include: heterogeneous, monolithic, hybrid integration etc.

Scalability: many manufacturing methods need to be adapted to large-scale production to meet the growing demand for photonic devices in the post Moore era. Current photonic manufacturing has relatively poor scalability compared to electronic ICs.

Materials innovation: the development of new materials with tailored optical properties will expand the design possibilities of photonic devices. Core semiconductor materials, besides the typical Silicon, are Gallium Arsenide, Lithium Niobate, Indium Phosphide, Silicon Nitride, and Transition-metal dichalcogenide (TMDCs).

Cost-Efficiency: reducing the cost of fabrication techniques, especially for lithography, 3D printing and self-assembly, is essential for widespread adoption of photonic devices. This could also call for cheaper fabrication tools and equipment.

Hybrid Platforms: combining different fabrication methods can offer unique advantages, leading to hybrid photonic platforms. Common semiconductor platforms include Complementary metal–oxide–semiconductor (CMOS), carbon systems including carbon nanotubes and graphene, and the latest magnetoelectric spin–orbit (MESO) [159] etc.

In conclusion, photonics has witnessed substantial advancements in fabrication platforms, enabling the creation of intricate and efficient photonic devices. Lithography, etching, self-assembly, and 3D printing have all played pivotal roles in shaping the future of photonics. With continued research and innovation, photonic devices are poised to revolutionize a wide range of applications, from

telecommunications and neuromorphics to medical imaging and AI, ushering in a new era of photonics on the nanoscale.

5 Recent Advances in Photonic Neuromorphic Computing

Figure 9 systematically compares biological (human brain), electronic, and photonic neuromorphic computing systems across the board. Key metrics / FoM being compared include: energy consumption, parallelism, latency, noise-level, computing speed, 3D or 2D topology, and learning capacity. From Figure 9 we see that photonic systems outperform electronic counterparts in all the aspects above and even match the energy consumption level of human brains. Similar trends can be observed in Figure 5.

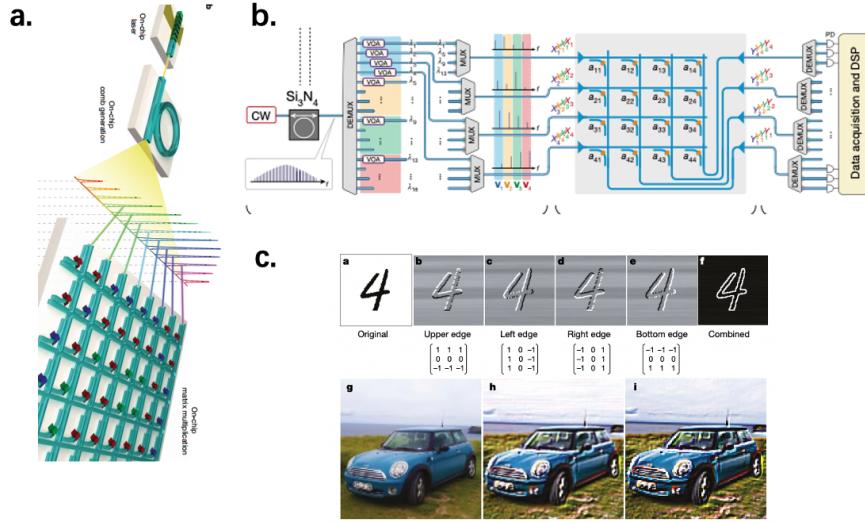


Figure 13: Convolutional neural networks (CNN) results demonstrated in the photonic tensor core paper [15]. a. Schematic of the photonic tensor core AI accelerator. b. Circuit diagram of the multiplexed all-optical convolutional tensor core. The input vectors are generated from a photonic frequency microcomb driven by a continuous-wave (CW) laser using wavelength division multiplexers. Right side involves the MAC unit (waveguides and PCMs). c. Convolution using sequential MVM operations on images of a handwritten digit 4 and a car. Using a high degree of parallelism, the processing time in c is reduced by a factor of 4. Reproduced with permission.[15] Copyright 2021, Springer Nature.

Table 1 more systematically tabulates and compares the recent milestone photonic neuromorphic paradigms for AI applications proposed by well-established institutions and groups around the globe. This table chronologically lists the technologies and methods used for achieving these representative photonic architectures, including the materials, devices, implementation platforms, encoding methods, and associated fabrication costs. Simultaneously, the performance of each architecture is reported as extracted from their individual papers, including the FoMs such as energy efficiency (TOP/J) and compute density ($TOP/mm^2/s$), where the former stands for tera (10^{12}) operations per joule while the latter tera operations per squared millimeter per second. The goal of state-of-the-art neuromorphic systems is achieving peta-level (10^{15}) in both energy efficiency and compute density. From Table 1 it's clear that while both energy efficiency and compute density have steadily risen with time, existing efforts struggle to meet the peta (10^{15}) threshold and they are still stuck on the tera-level in both FoMs for the most part. A comprehensive plot of these FoMs among various electronic and photonic neuromorphic systems is illustrated in Figure 5, which shows that photonic systems already outperform electronic counterparts in these FoMs. From the perspective of engineering, in Table 1 we also tabulate the estimated cost of fabrication for each work, aiming to compare and quantify how photonic neuromorphic systems' fabrication cost evolves with time or varies among different structures. The fabrication platform and devices used play a major role in the cost. Next, we observe that in the systems of [19] and [22], the input encoding (optical image signals) directly

Table 1: A summary of recent milestone photonic neuromorphic paradigms for AI and deep learning applications. Sources are listed chronologically. Only architectures that are CMOS-compatible chips are given compute density. Cost refers to the estimated fabrication cost per unit chip area according to the fabrication platform and devices used. [15] and [16], adopting similar technologies, were published on the same day. For pictorial illustration of each work, refer to Figure 3. NA=not available or not calculated. SOI=silicon-on-insulator.

Name abrv.	Technologies & methods	Energy efficiency (TOP/J)	Compute density (TOP/mm ² /s)	Est. cost /mm ²	Input encoding	Implementa platform	Reference
PNP	Mach-Zehnder interferometers, silicon photonics, photodiode, phase shifter	NA	NA	\$\$	Laser optical pulses	CMOS-compatible photonic chip	[14] 2017 Nature Photonics
D^2NN	3D printed lenses and optical diffraction	NA	NA	\$\$	Optical image signal	Free space & Benchtop	[13] 2018 Science
AONN	Spatial light modulator, Fourier lens, laser-cooled atom	NA	NA	\$\$	Optical image signal	Free space & Benchtop	[11] 2019 Optica
Spiking neurosynaptic network	Phase change material, micro-resonator, and wavelength division multiplexing	NA	NA	\$	Laser optical pulses	CMOS-compatible photonic chip	[21] 2019 Nature
Photonic tensor core	Phase change material, soliton microcombs, SiN micro-resonator, and wavelength division multiplexing	0.4	1.2	\$	Soliton frequency comb	CMOS-compatible photonic chip	[15] Jan 2021 Nature
Optical convolutional accelerator	Soliton microcombs, micro-resonator, Mach-Zehnder modulator, EDFA, and time-wavelength interleaving	1.27	8.061	\$\$	Electrical waveform	CMOS-compatible photonic chip	[16] Jan 2021 Nature
PDNN	PIN attenuator, SiGe photodiodes, grating coupler, and microring modulator	2.9	3.5	\$\$	Optical image signal	CMOS-compatible SOI Photonic chip	[19] 2022 Nature
PAIM	Meta-surface, optical diffraction, and FPGA	NA	NA	\$\$\$	Optical image signal	Free space & Benchtop	[18] 2022 Nature Electronics
VCSEL-ONN	VCSEL, diffractive optical element, and optical fanout	142.9	6	\$\$	Amplitude or phase of VCSEL	CMOS-compatible photonic chip	[17] 2023 Nature Photonics

comes from sensing targets, the energy cost and encoding rates of which are not taken into account, whereas in [15], [16], and [17], input encoding (e.g., frequency comb and electrical waveform) is a major difficulty to improve the overall system performance. Finally, we distinguish architectures that are integrated CMOS-compatible photonic chips from those that are free space bench-top implementations. Only those that are integrated chips are given the compute density (because it doesn't make sense to calculate those per-unit-area FoMs for free-space implementations) and we believe chip implementations offer much better computing and energy performances compared to free-space ones in the long run. As such, this table aims to summarize and contextualize existing approaches to photonic neuromorphic computing and provide the readers a bigger picture of the state-of-the-art development of this field up to now. Below, we introduce several of these representative works in detail.

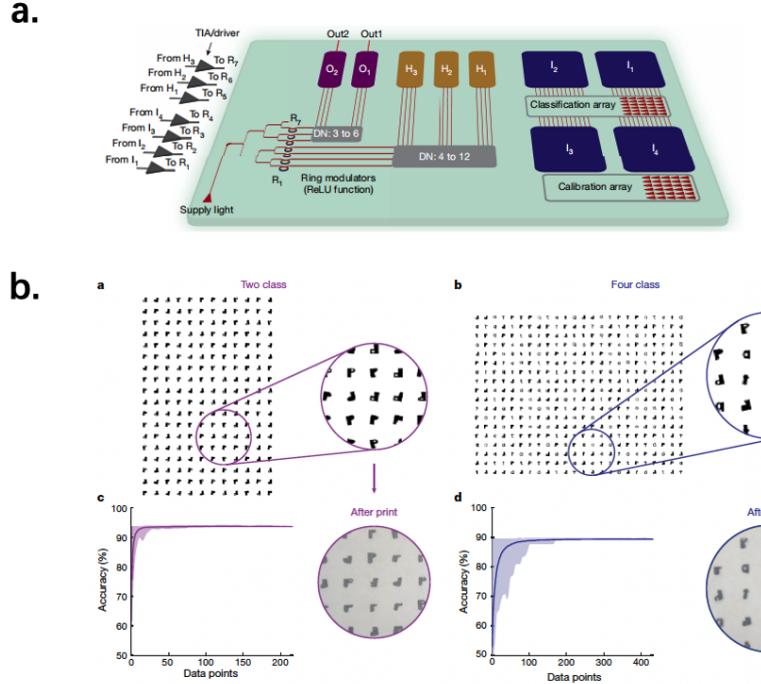


Figure 14: The implemented photonic classifier chip PDNN [19]. a. The top-level block diagram of PDNN and the structure of an implemented N-input photonic neuron, in which the weights of N optical input signals are adjusted using optical PIN attenuators and summed after photodetection using parallel PDs. An optical MRM realizes the ReLU non-linear activation function. b. Image classification demonstration, with the two-class dataset consisting of letters ‘p’ and ‘d’ (left) and four-class dataset consisting of ‘p’, ‘d’, ‘a’ and ‘t’ (right). Classification accuracy of both classes are plotted. Reproduced with permission.[19] Copyright 2022, Springer Nature.

In Figure 13, parallel, fast, and efficient convolution operations are realized by a photonic tensor core [15], which is enabled by technologies such as phase change material (PCM), soliton microcombs, SiN micro-resonator, and wavelength division multiplexing (WDM) etc. The tensor core can be likened to the optical equivalent of an ASIC, achieving highly parallelized photonic in-memory computing through PCM memory arrays and optical frequency combs based on photonic chips. The computation involves measuring the optical transmission of reconfigurable and non-resonant passive components, capable of operating at a bandwidth exceeding 14 GHz, constrained only by the speeds of modulators and photodetectors. In this study, the tensor core demonstrated a performance level of 2 tera-MAC operations per second (tera being a trillion). Notably, the convolutional operation, being a passive transmission measurement, theoretically allows calculations at the speed of light with very low power consumption (approximately 17 fJ per MAC), limited in experiments only by modulation and detection bandwidths. Recent advancements in hybrid integration, including soliton microcombs at microwave line rates, ultralow-loss Si₃N₄ waveguides, and high-speed on-chip

detectors and modulators, pave the way for the potential full CMOS wafer-scale integration of the photonic tensor core with silicon photonics in this innovative paradigm.

In Figure 14, the PDNN paper [19] demonstrated the first end-to-end PDNN photonic classifier chip that performs sub-nanosecond image classification through computation by propagation of optical waves, eliminating the need for an image sensor, digitization and large memory modules. PDNN processes optical waves that reach the on-chip pixel array as they traverse layers of neurons. Within each neuron, optical linear computation takes place, and the non-linear activation function is achieved opto-electronically, resulting in a classification time of less than 570 ps—comparable to a single clock cycle of cutting-edge digital platforms. Initially, a set of 500- μm -long PIN current-controlled attenuators is employed to individually tune the optical power in each input nanophotonic waveguide of the neuron. By forward biasing the PIN junction and injecting carriers, the power of the optical wave (i.e., the signal weight) for each neuron input can be adjusted. The outputs of attenuators are photodetected using SiGe photodiodes (PDs), and the resultant photocurrents are combined to generate the weighted sum of neuron inputs. A uniformly distributed supply light ensures a consistent per-neuron optical output range, facilitating scalability to large-scale PDNNs. Successful demonstrations include two-class and four-class classification of handwritten letters with accuracies exceeding 93.8% and 89.8%, respectively. Notably, the photonic classifier chip offers low energy consumption and ultra-low computation time, presenting revolutionary potential in applications like event-driven and salient object detection in autonomous driving. It can function as a stand-alone classifier or in conjunction with electronic processors, benefiting from the sub-nanosecond classification capability of the PDNN chip. Lastly, the PDNN chip is implemented in the silicon-on-insulator (SOI) process. SOI fabrication processes offer monolithic integration of electronic and photonic devices.

In Figure 15, the VCSEL-ONN paper [17] experimentally demonstrates a spatial-temporal-multiplexed ONN system that mitigates several key challenges faced by existing ONNs: low electro-optic conversion rate, large device footprint, and long latency. This paper explores neuron encoding using micrometer-scale vertical-cavity surface-emitting laser (VCSEL) arrays, characterized by efficient electro-optic conversion (< 5 attojoules per symbol) and a compact footprint (< 0.01 mm^2 per VCSEL array, not including the full setup with additional free-space components). Employing homodyne photoelectric multiplication enables matrix operations at the quantum-noise limit, accompanied by detection-based optical nonlinearity featuring an instantaneous response. The VCSEL-ONN architecture comprises N layers, where each layer performs a MVM followed by a nonlinear activation function, mimicking the biological neurons' "axon-synapse-dendrite" architecture. VCSEL-ONN encodes the input vector in i time steps to the amplitude or phase of a coherent laser oscillator (referred to as 'axon'), with its beam dendritically fanning out to j copies for parallel processing. With three-dimensional neural connectivity, this system achieves an energy efficiency of 7 fJ per operation and a compute density of $6 \text{ teraOP mm}^{-2}\text{s}^{-1}$, representing 100-fold and 20-fold improvements, respectively. Benchmarking on the MNIST dataset for digit classification yields an accuracy of 93.1% (over 98% of ground truth). The system is anticipated to perform MAC operations with an efficiency of 50 aJ/OP, primarily limited by memory access rather than optical energy consumption. It should also be noted that the full VCSEL-ONN includes some components in a free-space setup, which means that we are still some distance away from a fully integrated photonic neuromorphic system.

Although integrated photonics has been proven successful in photonic neural networks for neuromorphic computing, it has relatively poor scaling and the size of matrices remains small at the moment. Therefore, besides the above milestone integrated photonic chips, we introduce several recent free-space ONNs that enable nonlinearity operations.

Free-space optical implementations, in particular, hold promise for significant speed enhancements and reduced energy consumption in the deep learning realm, thanks to its inherent parallelism. However, realizing the essential nonlinear component of DNN in free-space optics presents challenges, limiting the platform's potential. Moreover, achieving parallel nonlinear activation for each data point adds complexity to preserving the advantages of linear free-space optics. To that extent, people have introduced a free-space optical neural network featuring diffraction-based linear weight summation and nonlinear activation facilitated by the saturable absorption of thermal atoms [160]. Specifically, they exploit the saturable absorption behavior of room-temperature rubidium atoms housed in a vapor cell and observed the nonlinearity in a single pass without any cavity, which allows point-by-point nonlinear activation of an incident image. In the paper, they demonstrated image classification of handwritten digits using only a single layer. Compared to a linear model, their optical nonlinearity

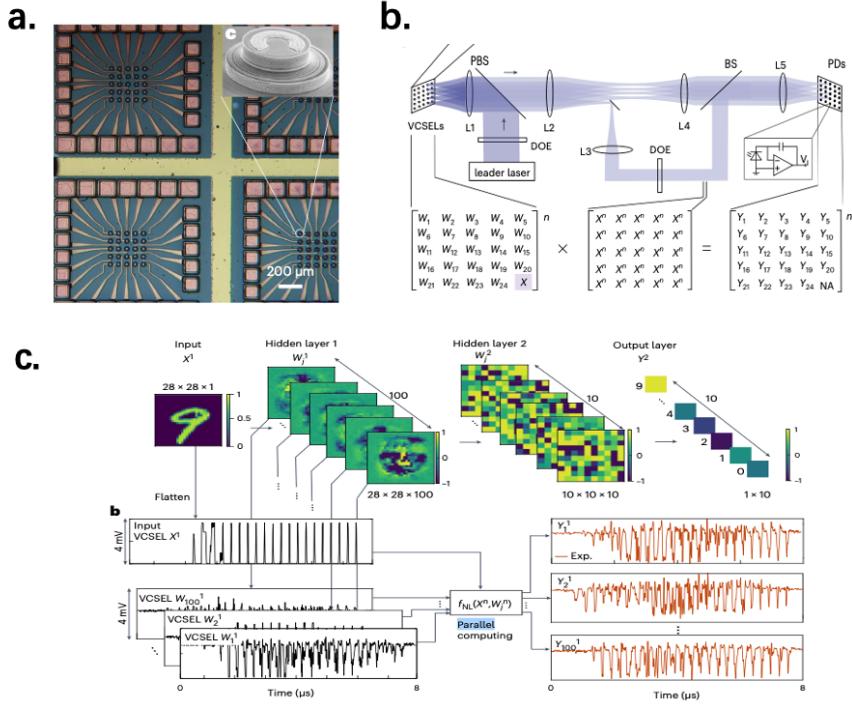


Figure 15: Parallel matrix-vector multiplication realized by VCSEL-ONN [17]. a. Fabricated VCSEL arrays. Arrays of 5×5 wire-bonded VCSELs on a GaAs substrate and an SEM image of a VCSEL emitter. b. Proposed architecture with 3D connectivity and photonic integration. In a 2D VCSEL array, the center unit is used as the axon and the others as weight VCSELs. The axon beam is fanned out to j copies, each overlapping a weighted beam onto a photodetector, generating photon currents corresponding to the homodyne product of the two laser fields. The input VCSEL is separated from the beam arrays using a beam magnifier (L1 and L2) and D-shaped mirror. c. Benchmarking of machine learning inference with VCSEL-ONN using MNIST. The input image in layer 1 is flattened and encoded in time steps to the phase of the X^n VCSEL. The weight matrix with 100 vectors is encoded to 100 individual weighting VCSELs. Parallel multiplication results in MVM from 100 readout channels. Reproduced with permission.[17] Copyright 2023, Springer Nature.

contributes to a 6% improvement in classification accuracy [160]. This platform maintains the significant parallelism inherent in free-space optics, even with physical nonlinearities, paving the way for the broader adoption of ONNs in neuromorphic computing. Another work [11] utilized a similar approach to this one.

In the realm of image sensing, the process of obtaining an object's location or shape involves analyzing captured images in digital computers. A novel approach to image sensing involves using optical systems not for traditional imaging but as encoders that compress images optically into low-dimensional spaces by extracting key features. However, the effectiveness of these encoders is typically constrained by the linearity. In a recent study [161], researchers introduced a free-space, nonlinear, and multilayer ONN encoder for image sensing. This ONN encoder utilizes an image intensifier as an optical-to-optical nonlinear activation (OONA) function. The ONN encoder comprises an optical matrix–vector multiplier unit, an OONA (nonlinear) unit, a second optical matrix–vector multiplier, and a camera. Compared to similarly sized linear optical encoders, this nonlinear ONN demonstrates superior performance across various tasks, including computer vision, flow-cytometry image classification, and identifying objects in a 3D-printed real scene. Particularly for computer vision tasks employing incoherent broadband illumination, their paradigm allows for significant reductions in camera resolution requirements and electronic post-processing complexity. Overall, employing free-space ONNs holds promise for image-sensing applications that can operate accurately with fewer pixels, fewer photons, lower power consumption, increased throughput, and reduced latency.

6 Emerging Areas in Photonics for Neuromorphic Computing

6.1 Emerging light sources and their potential impact

Emerging nanophotonic technologies are leading the way in scientific and engineering progress, offering innovative approaches to observe the world through precise control of light at the nanoscale. In the wake of the seminal discovery of the quantum Hall effect, the physics community has embarked on a journey marked by significant advancements in the realm of topological insulators (TI). For instance, in the presence of a magnetic field within the microwave frequency range, researchers observed the unidirectional transmission of electromagnetic fields, revealing their remarkable anti-scattering characteristics and robustness [162]. This unique attribute renders topological insulators exceptionally suitable for its incorporation in next-gen nanophotonic devices. Simply put, TI enhances the fault tolerance and robustness of nano-scale on-chip integrated optical systems, while enabling functionalities that were previously challenging to attain with conventional optical components.

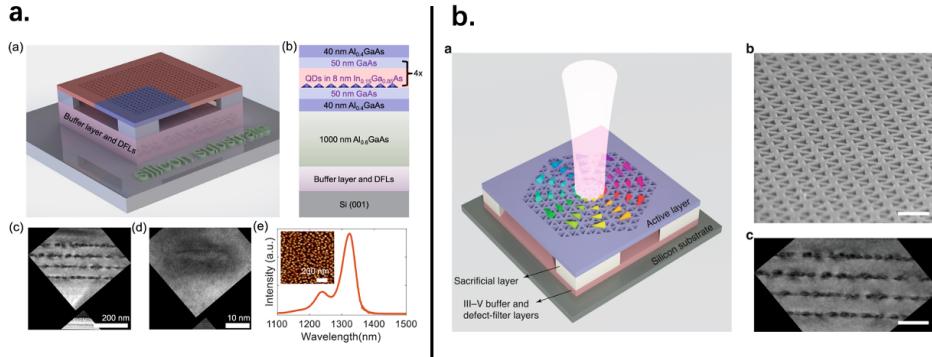


Figure 16: State-of-the-art topological photonic lasers. a. Schematic structure of the fabricated topological corner state nanolasers monolithically integrated on a CMOS-compatible silicon substrate [163]. High-resolution TEM images of the four stacked InAs/GaAs QDs and a single QD in the active region, respectively, are shown. Also shows the measured PL spectra of as-grown InAs/GaAs QDs on silicon. b. Conceptual illustration of a topological Dirac-vortex microcavity laser fabricated on a silicon substrate [115]. The photonic crystal structure was defined in the active layer and suspended by partially removing the sacrificial layer. The III–V buffer and defect-filter layers were carefully optimized to minimize the effects of lattice mismatch between the III–V materials and silicon substrate. Tilted-view SEM image of the fabricated topological laser and cross-sectional bright-field TEM image of the active layer containing four-stack InAs/InGaAs QD layers are shown. Panels reproduced with permission from: a. Reproduced with permission.[163] Copyright 2022, American Chemical Society. b. Reproduced under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>).[115] Copyright 2023, the Authors and Springer Nature.

Recently, the emergence of topological photonics marks a profound departure from conventional optical paradigms, introducing the concept of "topological invariants" and "topologically protected state" to the domain of optics. This innovative approach delves into the intriguing premise that wave functions of electromagnetic fields remain impervious to alteration amidst fluctuating geometric structures. Consequently, topological photonic devices began to widely emerge, distinguished by their exceptional optical performance attributes and immunity to perturbations. These include an inherent robustness, a remarkable capacity to resist interference, the effective suppression of photon backscattering, and the realization of a notably high free spectral range (FSR). All of these above attributes render topological photonics an attractive candidate for building integrated on-chip light sources for neuromorphic computing, such as those lasers employed in [14], [21], [15], and especially [17].

In non-Hermitian systems, nanoscale topological photonic crystals, due to their remarkable robustness and anti-scattering properties, show great potential in laser technology. By harnessing the unique characteristics of boundary states, bulk states, corner states (Figure 16a), and other features inherent to topological photonic crystals, light sources with diverse optical properties can be realized [163–166].

The photonic band structure at the Dirac point brings distinct property of zero refractive index, and combined with the transmission property of topological photonic crystals, a kind of vortex state laser can be designed by precisely controlling the phase of the electromagnetic field [115, 167] (Figure 16b). This breed of laser, allowing for meticulous control of the light field distribution at the nanoscale, boasts an impressively high free spectral range and surface emission characteristics, thus emerging as a highly competitive light source for neuromorphic computing. This novel Dirac-vortex laser [115] has the potential to outperform the VCSEL lasers utilized in [17] in terms of computing speed, energy efficiency, electro-optic conversion rate, and stability. In addition, to facilitate the transition of topological photonics into practical applications, researchers conceived topological quantum cascade lasers powered by terahertz waves, thereby significantly amplifying the potential of topological lasers for neuromorphic computing [168, 169].

Within Hermitian topological systems, valley photonic crystals were used to achieve lossless topological photonic insulator waveguides, even when subjected to sharp bends [170]. Furthermore, researchers delved into the exceptional properties of topological slow-light waveguides [171] and polarization beam splitters [172], all of which can be integrated in neuromorphic computing chips in conjunction with the aforementioned topological lasers. Importantly, the topological protection after laser emission will still be held as long as the coupling components (such as waveguides) belong to the same class of topological photonic structure as the laser. Topological photonic components also can be integrated into a CMOS-compatible silicon-on-insulator (SOI) platform [173], exemplifying the immense potential of integrated topological photonics within the field of neuromorphic computing.

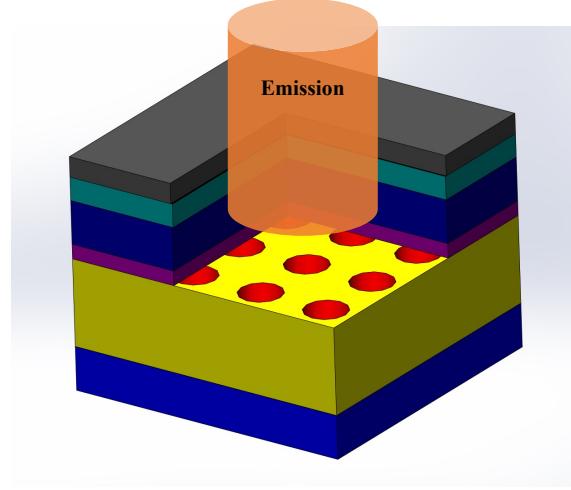


Figure 17: Schematic of a standard PCSEL device with circular air holes in the PhC layer. Multiple epitaxial layers and substrates are symbolically shown, with their names omitted for simplicity. Laser beam is emitted vertically.

In addition to topological lasers, Photonic Crystal Surface Emitting Lasers (PCSELs) [103, 104, 174, 175] (Figure 17) are a novel type of laser that combines the advantages of photonic crystals (PhCs) [176] and Vertical Cavity Surface Emitting Lasers (VCSELs) [177]. PhCs are artificially engineered structures with periodic changes in refractive index in one, two, or three dimensions, creating a bandgap that restricts the propagation of light in specific frequency ranges. VCSELs are lasers that emit light perpendicular to the surface of a semiconductor, facilitating efficient coupling with optical fibers and other optical components. PCSELs merge these technologies to create lasers with multiple benefits compared to traditional lasers, offering the best of both worlds.

The fundamental structure of a PCSEL includes a PhC layer, an active layer, additional cladding layers, substrates, p-n junctions, and electrodes at the ends. The PhC layer mainly functions as a resonant cavity, while the active layer, typically composed of III-V materials (such as InP/InGaP, GaAs/InGaAs/AlGaAs, GaN/InGaN, etc.), is positioned in the middle of the PCSEL to induce laser emission when a population inversion of charge carriers is achieved upon reaching a certain threshold

[104, 178]. Population inversion occurs when there are more electrons in higher energy states than in lower energy states, allowing for stimulated emission of photons when an electrical current is applied. When an electrical pumping current is introduced into the active layer material, it emits laser light that is effectively confined and amplified within the resonance cavity. Furthermore, the active layer can contain quantum dots or quantum wells, which increase the recombination rate of spontaneous emission, significantly enhancing the lasing effect. Consequently, PCSELs have superior characteristics than conventional VCSELs.

6.2 Emerging silicon-on-insulator (SOI) paradigm and its potential impact

Monolithically integrated microcavity lasers on heterogeneous epitaxy-enabled SOI substrates (Figure 17) is a promising method to achieve in-plane light source coupling within photonic integrated circuits (PIC). The SOI paradigm can greatly improve the integration density and scalability while reducing the cost of PICs, while at the same time offering monolithic integration of electronic and photonic devices on the same chip. Besides, it can help overcome the common large lattice mismatch, large diffusivity of indium adatoms, and the thick buffer layer issues of III-V materials monolithically grown on Si substrate. In short, SOI is expected to play a significant role in light source integration of PICs for short-distance optical communication and data centers. Faster speed, ultrasmall footprint, low power consumption, low cost, and CMOS compatible fabrication ensure the significant position of SOI-based PICs in data communication network and optical computing.

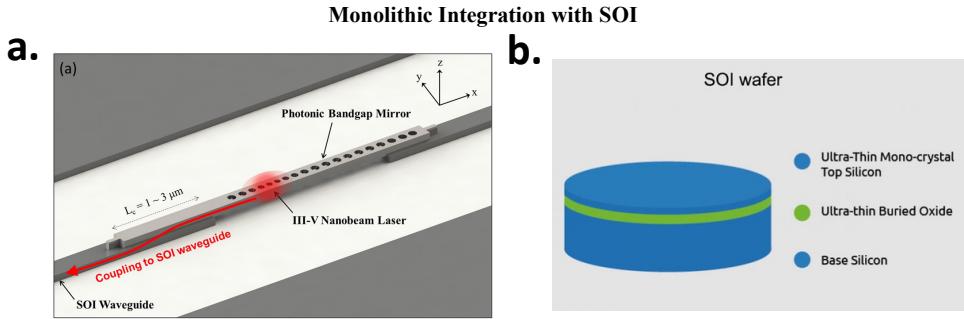


Figure 18: a. Schematic diagram of a PhC nanobeam laser grown on silicon-on-insulator (SOI) substrate, monolithically integrated with Si waveguides, electrodes, and other on-chip devices. Reproduced with permission.[179] Copyright 2017, American Chemical Society. b. side view of a typical SOI wafer, where the buried oxide is Silica.

6.3 Emerging optical encoder-ANNs and their potential impact

The recent advancements in utilizing light for conducting large-scale linear operations in parallel have sparked numerous demonstrations of optics-integrated artificial neural networks (ANNs). Nevertheless, a definitive system-level superiority of optics over entirely digital ANNs has yet to be firmly established. While optical systems excel in efficiently executing linear operations, the absence of nonlinearity and signal regeneration necessitates high-power and low-latency signal transmission between optical and electronic systems. Moreover, substantial power requirements for lasers and photodetectors, often overlooked in energy consumption calculations, further complicate the picture.

In this context, instead of merely translating conventional digital operations into optics, people have developed a hybrid optical-digital ANN through utilizing a meta-optical encoder [180]. This hybrid model operates using incoherent light, making it compatible with ambient light conditions. By maintaining consistent latency and power levels between a fully digital ANN and our hybrid optical-digital ANN, they have identified a regime characterized by low power and latency. Within this regime, an optical encoder outperforms a purely digital ANN in terms of classification accuracy on MNIST dataset. Their estimates suggest that the optical encoder facilitates operation rates of over 10 kHz for a hybrid ANN, consuming only 23 mW of power. However, within this regime, the overall classification accuracy is slightly lower compared to what can be achieved with higher power and latency settings.

These findings suggest that optics can offer advantages over digital ANNs in scenarios where prioritizing lower power consumption and latency is paramount, even if it entails some compromise on overall performance [180].

7 Remaining Challenges and Future Directions

7.1 Strategies for improving device performance, scalability, and reliability

Table 2: Strategies for improving photonic device performance, scalability, and reliability. The realm of photonics plays a vital role in improving data transfer speeds in optical communication, thus achieving scalability in photonic components is a pressing goal. Device reliability is paramount, particularly in critical applications like aerospace and healthcare. Strategies to enhance reliability encompass both design and materials.

Performance	Scalability	Reliability
Advanced Transistor Architectures: The introduction of novel transistor architectures, such as FinFETs and gate-all-around (GAA) nanosheets, has significantly improved the performance of microprocessors. These designs offer enhanced control of the current flow, reduced leakage, and higher switching speeds [181].	Photonic Integrated Circuits (PICs): PICs have emerged as a key strategy to integrate multiple optical modules onto a single chip, reducing size and increasing functionality [182]. Advances in PIC design and fabrication have improved the performance of devices like optical transceivers and lasers.	Redundancy and Error Correction: In electronic systems, redundancy and error correction techniques are employed to improve reliability. These methods ensure that even in the presence of defects or failures, the system continues to function correctly [183].
High-Mobility Materials: The integration of high-mobility materials, like gallium nitride (GaN) and indium gallium arsenide (InGaAs), has led to faster and more efficient optoelectronic devices. GaN power devices, for example, have found applications in power electronics and RF amplifiers [184].	Meta-surfaces: Meta-surfaces consist of subwavelength nanostructures that manipulate light in novel ways. These structures are integral in creating flat optical components, enabling highly compact and scalable photonic devices [185, 186].	Photonic Error Correction: There have been experimental demonstrations of deterministic real time training and error correction in integrated photonic systems [187–189] to realize efficient deep learning in photonic neural networks (via in situ training).
Quantum Devices: Quantum devices, such as quantum-dot transistors and quantum cascade lasers, offer unique properties that can revolutionize electronics and photonics. Quantum computing, for example, has the potential to solve complex problems at unprecedented speeds [190, 191].	Additive Manufacturing: Additive manufacturing, including 3D printing, allows for the rapid and cost-effective production of complex components. It has applications in aerospace, healthcare, and customized electronic packaging [192, 193]; Monolithic 3D Integration: Monolithic 3D integration is a disruptive manufacturing technique that enables the stacking of active devices in the vertical direction (also known as epitaxy), enhancing performance and reducing interconnect lengths [194–197].	Reliability Testing: Rigorous reliability testing, including accelerated life tests, is conducted to identify and mitigate potential failure modes. This process ensures that devices meet their specified lifetime and performance criteria. [198].

As neuromorphic technology continues to advance and demands are getting higher, the performance, scalability, and reliability of photonic devices become increasingly critical. Achieving higher efficiency, speed, and functionality is essential to meet the ever-growing demands of modern computing tasks. This section (Table 2) will address recent developments and strategies implemented to enhance device performance, scalability, and reliability in neuromorphic photonics, where engineers and scientists are exploring a multitude of strategies. These encompass device engineering techniques, the incorporation of novel materials, and advancements in manufacturing processes. These strategies are integral in fields as diverse as semiconductor optoelectronics, photonics, and integrated systems. For instance, [188] proposed a reconfigurable integrated photonic processor that performs in situ training of vowel recognition with high accuracy, whose measurement results for the output light

power are monitored in a real-time manner and feedback from the detection is delivered to the encoder to update the pumping pattern for self-error correction. Another work [187] proposed Silicon photonic architecture that employs the direct feedback alignment training algorithm, which trains neural networks using error feedback rather than error backpropagation, and can operate at speeds of trillions of multiply–accumulate (MAC) operations per second while consuming less than one picojoule per MAC operation. Their system includes MRRs that modulate the incoming laser light with the error vector e and transimpedance amplifiers (TIAs) with tunable gain to convert photocurrent to voltage and scale it to implement the Hadamard product. The authors believe that in the quest for better performance, scalability, and reliability, manufacturing processes and fabrication techniques play a pivotal role and should be the focus of neuromorphic and semiconductor researchers now and in the immediate future.

7.2 PIC's bottlenecks, remaining challenges, and future directions of neuromorphic photonics

Left panel of Table 3 first compares PICs to electronic ICs and lists the areas where electronic still dominates photonic systems. It then lists some remaining challenges to be solved and future directions to explore and improve for the PIC in the right panel. From this comparison, we can conclude that despite excelling at areas such as energy consumption, parallelism, and latency, PICs still largely lag behind electronic ICs in many other aspects, most evidently cost, scalability, integratability and footprint. The fact that there are still bulky bench-top implementations in Table 1 is testament that PICs have poor integratability and large footprint. As a matter of fact, most recent state-of-the-art PICs are at the state of ICs almost 60 years ago, when Intel produced the very first batch of computer chips. As a result, we are still decades away from seeing photonic chips actually deployed in our smart phones, laptops, and iPads, and even farther away from having full-blown photonic computers capable of running large AI models such as chatGPT or AlphaGo or solving complex scientific problems such as molecular simulation or finite-element analysis. In other words, the total computing power of PICs (capable of handling few-layer DNNs and less than 10,000s of parameters) at the moment is simply not a rival for electronic ICs (capable of handling DNNs with 100s of layers and trillions of parameters). On the flip side, however, we already have some number of PICs deployed in large-scale data centers and telecommunication systems where integratability and footprint are not of concern. Another major concern of PICs is its cost. As seen in Table 1, the cost to fabricate a photonic chip can be quite high, almost three times more expensive than standard ICs. As a matter of act, cost has always been a critical challenge of silicon photonics that inhibits its widespread adoption in the IT and computing industry. So the industry should focus on reducing the cost by either inventing new materials, revolutionizing the fabrication/integration technologies, or simply expanding scalability. Improved scale and scalability will effectively drive the lowering of cost just as the IC industry has experienced over the past 50 years. Furthermore, latest monolithic integration and SOI techniques could largely improve the integration density and bring down cost. All in all, as the Moore's law comes to an end and the photonic version of the Moore's law begins to take off, we expect to see a considerable improvement in PIC's cost, scalability, integratability, and total computing capacity. PICs will eventually co-exist, if not replace, ICs as the backbone of future computing systems.

8 Conclusion

Advances in photonics have catalyzed a transformation in computational technologies, with the integration of optoelectronics onto photonic platforms leading the charge. This integration has facilitated the emergence of PICs, which act as the building blocks for ultra-fast artificial neural networks and are pivotal in the creation of next-generation computational devices. These devices are engineered to address the intensive computational demands of machine learning and AI applications across sectors including healthcare diagnostics, complex language processing, telecommunications, high-performance computing, and immersive virtual environments.

Despite the advancements, conventional electronic systems exhibit limitations in speed, signal interference, and energy efficiency. Neuromorphic photonics, characterized by its ultra-low latency, emerges as a groundbreaking solution, carving out a new trajectory for the advancement of AI and ONNs. This review casts a spotlight on the latest developments in neuromorphic photonic systems from the perspective of photonic engineering and material science, critically analyzing the emergent and anticipated challenges, and mapping out the scientific and technological innovations necessary to surmount these obstacles.

Table 3: Left panel: photonic integrated circuits compared to electronic integrated circuits, in terms of cost, scalability, integratability, footprint, computing capacity (throughput) etc.. Right panel: remaining challenges and future directions for photonic devices' research.

Electronic IC	PIC	Challenges and Future Directions of PIC
Cost <i>low</i> (mature fabrication processes and large-scale production)	Cost <i>medium</i> (mature fabrication processes but lack of scale)	Power Efficiency: Enhancing power efficiency remains a significant challenge. Reducing power consumption is crucial, particularly in mobile devices and data centers.
Scalability <i>excellent</i>	Scalability <i>good</i> but yet to be demonstrated	Quantum Limitations: While quantum devices show immense promise, they also present challenges such as decoherence and error correction.
Integratability <i>excellent</i> (billions of transistors per chip, but coming to an end)	Integratability <i>poor</i> (1000s of devices per chip, and continuing to grow)	Sustainability: The electronics industry faces questions of sustainability and responsible sourcing of materials, which are central to achieving carbon neutral and tackling the climate change. The photonics industry should avoid the same pattern by reducing carbon footprint.
Footprint <i>small</i> (nanometer-scale)	Footprint <i>large</i> (micro-millimeter scale)	Security: As devices become more connected, ensuring the security and privacy of data remains a critical concern [199].
Total computing capacity <i>high</i> (DNN w/ hundreds of layers & up to trillions of parameters)	Total computing capacity <i>low</i> (few-layer DNN & up to 10,000s of parameters)	Other directions: photonics-electronics co-packaging, reducing cost of electron-beam lithography, addressing time-consuming FEA or FDTD simulations...

The focus is on an array of neuromorphic photonic AI accelerators, examining the spectrum from classical optics to sophisticated PIC designs. It scrutinizes their operational efficiency, particularly in terms of operations per watt, through a detailed comparative analysis emphasizing key technical parameters. The discussion pivots to specialized technologies such as VCSEL/PCSEL and frequency microcomb-based accelerators, accentuating the latest innovations in photonic modulation and wavelength division multiplexing for effective neural network training and inference.

Acknowledging the current technological barriers in achieving computational efficiencies at the PetaOPs/Watt threshold, the review explores prospective strategies to enhance these critical performance metrics. These include the emerging topological insulators and PCSELS as well as strategies to advance fabrication, system scalability and reliability. The exploration aims to not only chart the current landscape but also to forecast the trajectory of neuromorphic photonics in pushing the frontiers of AI capabilities in the near future. All in all, as the Moore's law comes to an end and the photonic version of the Moore's law begins to take off, we expect to see a considerable improvement in PIC's cost, scalability, integratability, and total computing capacity. PICs will eventually replace ICs as the backbone of future computing systems.

Acknowledgement

This work is supported by National Natural Science Foundation of China (NSFC) under Grant No.62174144, Shenzhen Science and Technology Program under Grant No.JCYJ20210324115605016, No.JCYJ20210324120204011, No.JSGG20210802153540017, No.KJZD20230923115114027, and No.JCYJ20220818102214030, Guangdong Key Laboratory of Optoelectronic Materials and Chips under Grant No.2022KSYS014, Shenzhen Key Laboratory Project under Grant No.ZDSYS201603311644527; Longgang Key Laboratory Project under Grant No.ZSYS2017003 and No.LGKCZSYS2018000015; Shenzhen Research Institute of Big Data; Innovation Program for Quantum Science and Technology under Grant No.2021ZD0300701. The authors'd like to thanks Mr. Ceyao Zhang and Dr. Feng Yin for their fruitful discussion on ML and LLMs, and thank Mr. Qi Xin, Pengyu Zhou and Mr. Ping Sun for their fantastic art drawings and CAD modellings, and thank Dr. Hongjie Wang and Mr. Xiaolei Shen for their mentorship and revision tips.

Conflict of interests

The authors declare no conflict of interests.

Author contributions

Z.Z conceived and proposed the writing project. R.L. and Z.Z. designed the structure and outline of the paper. R.L. led and arranged the writing of the whole paper. R.L., Y.G., H.H., Y.Z., S.M. wrote the paper together. S.M. and Y.Z. applied for copyright permissions. Z.Z. supervised and mentored the project. Z.Z. funded the project.

References

- [1] S. Russell and P. Norvig. 3rd ed. Prentice Hall, 2010.
- [2] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al. In: *arXiv preprint arXiv:2206.07682* (2022).
- [3] J. Tang, F. Yuan, X. Shen, Z. Wang, M. Rao, Y. He, Y. Sun, X. Li, W. Zhang, Y. Li, et al. In: *Advanced Materials* 31.49 (2019), p. 1902761.
- [4] N. K. Upadhyay, H. Jiang, Z. Wang, S. Asapu, Q. Xia, and J. Joshua Yang. In: *Advanced Materials Technologies* 4.4 (2019), p. 1800589.
- [5] F. Torres, A. C. Basaran, and I. K. Schuller. In: *Advanced Materials* (2023), p. 2205098.
- [6] C. D. Schuman, S. R. Kulkarni, M. Parsa, J. P. Mitchell, P. Date, and B. Kay. In: *Nature Computational Science* 2.1 (2022), pp. 10–19.
- [7] B. J. Shastri, A. N. Tait, T. Ferreira de Lima, W. H. Pernice, H. Bhaskaran, C. D. Wright, and P. R. Prucnal. In: *Nature Photonics* 15.2 (2021), pp. 102–114.
- [8] I. A. Williamson, T. W. Hughes, M. Minkov, B. Bartlett, S. Pai, and S. Fan. In: *IEEE Journal of Selected Topics in Quantum Electronics* 26.1 (2019), pp. 1–12.
- [9] M. Miscuglio, A. Mehrabian, Z. Hu, S. I. Azzam, J. K. George, A. V. Kildishev, M. Pelton, and V. J. Sorger. In: *Optical Materials Express* 8.12 (Dec. 2018). DOI: 10.1364/ome.8.003851MAG ID: 2900709571, pp. 3851–3863.
- [10] G. Mourigas-Alexandris, A. Tsakyridis, N. Passalis, A. Tefas, A. Tefas, K. Vyrskinos, and N. Pleros. In: *Optics Express* 27.7 (Apr. 2019). DOI: 10.1364/oe.27.009620MAG ID: 2923894441, pp. 9620–9630.
- [11] Y. Zuo, B. Li, Y. Zhao, Y. Jiang, Y.-C. Chen, P. Chen, G.-B. Jo, J. Liu, and S. Du. en. In: *Optica* 6.9 (Sept. 2019), p. 1132. ISSN: 2334-2536.
- [12] T. W. Hughes, M. Minkov, Y. Shi, and S. Fan. In: *Optica* 5.7 (2018), pp. 864–871.
- [13] X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, and A. Ozcan. In: *Science* 361.6406 (Sept. 2018), pp. 1004–1008.
- [14] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Laroche, D. Englund, and M. Soljačić. en. In: *Nature Photonics* 11.77 (July 2017), pp. 441–446. ISSN: 1749-4893.
- [15] J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, M. Stappers, M. Le Gallo, X. Fu, A. Lukashchuk, A. S. Raja, et al. In: *Nature* 589.7840 (2021), pp. 52–58.

- [16] X. Xu, M. Tan, B. Corcoran, J. Wu, A. Boes, T. G. Nguyen, S. T. Chu, B. E. Little, D. G. Hicks, R. Morandotti, et al. In: *Nature* 589.7840 (2021), pp. 44–51.
- [17] Z. Chen, A. Sludds, R. Davis III, I. Christen, L. Bernstein, L. Ateshian, T. Heuser, N. Heermeier, J. A. Lott, S. Reitzenstein, et al. In: *Nature Photonics* 17.8 (2023), pp. 723–730.
- [18] C. Liu, Q. Ma, Z. J. Luo, Q. R. Hong, Q. Xiao, H. C. Zhang, L. Miao, W. M. Yu, Q. Cheng, L. Li, et al. In: *Nature Electronics* 5.2 (2022), pp. 113–122.
- [19] F. Ashtiani, A. J. Geers, and F. Aflatouni. In: *Nature* 606.7914 (2022), pp. 501–506.
- [20] D. Psaltis, D. Brady, X.-G. Gu, and S. Lin. In: *Nature* 343.6256 (1990), pp. 325–330.
- [21] J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. H. Pernice. In: *Nature* 569.7755 (2019), pp. 208–214.
- [22] Y. Chen, M. Nazhamaiti, H. Xu, Y. Meng, T. Zhou, G. Li, J. Fan, Q. Wei, J. Wu, F. Qiao, et al. In: *Nature* (2023), pp. 1–10.
- [23] M. A. Zidan, J. P. Strachan, and W. D. Lu. In: *Nature electronics* 1.1 (2018), pp. 22–29.
- [24] X. Huang, C. Liu, Y.-G. Jiang, and P. Zhou. In: *Chinese Physics B* 29.7 (2020), p. 078504.
- [25] S. Agwa and T. Prodromakis. In: *Frontiers in Nanotechnology* 5 (), p. 1147396.
- [26] C. E. Leiserson, N. C. Thompson, J. S. Emer, B. C. Kuszmaul, B. W. Lampson, D. Sanchez, and T. B. Schardl. In: *Science* 368.6495 (2020), eaam9744.
- [27] N. S. Kim, D. Chen, J. Xiong, and W. H. Wen-me. In: *IEEE Micro* 37.4 (2017), pp. 10–18.
- [28] S. Xue. In: *International Core Journal of Engineering* 7.8 (2021), pp. 330–334.
- [29] M. A. Nahmias, T. F. De Lima, A. N. Tait, H.-T. Peng, B. J. Shastri, and P. R. Prucnal. In: *IEEE Journal of Selected Topics in Quantum Electronics* 26.1 (2019), pp. 1–18.
- [30] J. Backus. In: *Communications of the ACM* 21.8 (1978), pp. 613–641.
- [31] I. Arikpo, F. Ogban, and I. Eteng. In: *Global Journal of Mathematical Sciences* 6.2 (2007), pp. 97–103.
- [32] O. Mutlu, S. Ghose, J. Gómez-Luna, and R. Ausavarungnirun. In: *Microprocessors and Microsystems* 67 (2019), pp. 28–41.
- [33] K. Olukotun and L. Hammond. In: *Queue* 3.7 (2005), pp. 26–29.
- [34] C. Cheng, P. J. Tiw, Y. Cai, X. Yan, Y. Yang, and R. Huang. In: *Science China Information Sciences* 64 (2021), pp. 1–46.
- [35] F. Yazdanpanah, C. Alvarez-Martinez, D. Jimenez-Gonzalez, and Y. Etsion. In: *IEEE Transactions on Parallel and Distributed Systems* 25.6 (2013), pp. 1489–1509.
- [36] M. Shaafiee, R. Logeswaran, and A. Seddon. In: *2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence*. IEEE. 2017, pp. 199–203.
- [37] T. Wang, S.-Y. Ma, L. G. Wright, T. Onodera, B. C. Richard, and P. L. McMahon. In: *Nature Communications* 13.1 (2022), p. 123.
- [38] R. Hamerly, L. Bernstein, A. Sludds, M. Soljačić, and D. Englund. In: *Physical Review X* 9.2 (2019), p. 021032.
- [39] C. Neagu. PhD thesis. Politecnico di Torino, 2023.
- [40] C. Mead and M. Ismail. Vol. 80. Springer Science & Business Media, 1989.
- [41] W. A. Wulf and S. A. McKee. In: *ACM SIGARCH computer architecture news* 23.1 (1995), pp. 20–24.
- [42] M. Horowitz. In: *2014 IEEE international solid-state circuits conference digest of technical papers (ISSCC)*. IEEE. 2014, pp. 10–14.
- [43] D. Ielmini and H.-S. P. Wong. In: *Nature electronics* 1.6 (2018), pp. 333–343.
- [44] G. Chirkov and D. Wentzlaff. In: *Proceedings of the 37th International Conference on Supercomputing*. 2023, pp. 410–422.
- [45] T. N. Theis and H.-S. P. Wong. In: *Computing in Science & Engineering* 19.2 (2017), pp. 41–50.
- [46] H. N. Khan, D. A. Hounshell, and E. R. Fuchs. In: *Nature Electronics* 1.1 (2018), pp. 14–21.
- [47] L. B. Kish. In: *Physics Letters A* 305.3 (2002), pp. 144–149. ISSN: 0375-9601.
- [48] T. Zanotti, F. M. Puglisi, and P. Pavan. In: *IEEE Journal of the Electron Devices Society* 8 (2020), pp. 757–764.
- [49] S. A. McKee. In: *Proceedings of the 1st conference on Computing frontiers*. 2004, p. 162.
- [50] A. Saulsbury, F. Pong, and A. Nowatzky. In: *ACM SIGARCH Computer Architecture News* 24.2 (1996), pp. 90–101.
- [51] S. M. Goodnick and J. Bird. In: *IEEE Transactions on Nanotechnology* 2.4 (2003), pp. 368–385.
- [52] B. Gopi, J. Logeshwaran, J. Gowri, and V. Aravindarajan. In: *NeuroQuantology* 20.8 (2022), pp. 5999–6010.
- [53] X. Zhang, A. Huang, Q. Hu, Z. Xiao, and P. K. Chu. In: *physica status solidi (a)* 215.13 (2018), p. 1700875.

- [54] N. Qiao, H. Mostafa, F. Corradi, M. Osswald, F. Stefanini, D. Sumislawska, and G. Indiveri. In: *Frontiers in neuroscience* 9 (2015), p. 141.
- [55] B. V. Benjamin, P. Gao, E. McQuinn, S. Choudhary, A. R. Chandrasekaran, J.-M. Bussat, R. Alvarez-Icaza, J. V. Arthur, P. A. Merolla, and K. Boahen. In: *Proceedings of the IEEE* 102.5 (2014), pp. 699–716.
- [56] S. G. Ramasubramanian, R. Venkatesan, M. Sharad, K. Roy, and A. Raghunathan. In: *Proceedings of the 2014 international symposium on Low power electronics and design*. 2014, pp. 15–20.
- [57] E. Farquhar, C. Gordon, and P. Hasler. In: *2006 IEEE international symposium on circuits and systems*. IEEE. 2006, 4–pp.
- [58] A. Grubl, S. Billaudelle, B. Cramer, V. Karasenko, and J. Schemmel. In: *Journal of Signal Processing Systems* 92 (2020), pp. 1277–1292.
- [59] D. Kuzum, S. Yu, and H.-S. P. Wong. In: *Nanotechnology* 24.38 (Sept. 2013), p. 382001.
- [60] C. S. Sherrington. In: *Scientific and Medical Knowledge Production, 1796-1918*. Routledge, 2023, pp. 217–253.
- [61] E. R. Kandel, J. H. Schwartz, T. M. Jessell, S. Siegelbaum, A. J. Hudspeth, S. Mack, et al. Vol. 4. McGraw-hill New York, 2000.
- [62] C. Mead. In: *Proceedings of the IEEE* 78.10 (1990), pp. 1629–1636.
- [63] S. B. Furber, F. Galluppi, S. Temple, and L. A. Plana. In: *Proceedings of the IEEE* 102.5 (2014), pp. 652–665.
- [64] K. Meier. Tech. rep. FP6-2004-IST-FET Proactive, Part B. Kirchhoff Institut f"ur Physik, Ruprecht . . ., 2004.
- [65] T. Hylton. In: *DARPA SYNAPSE Bidder's Workshop and Teaming Meeting*. 2008.
- [66] Q. Xia, W. Robinett, M. W. Cumbie, N. Banerjee, T. J. Cardinali, J. J. Yang, W. Wu, X. Li, W. M. Tong, D. B. Strukov, et al. In: *Nano letters* 9.10 (2009), pp. 3640–3645.
- [67] A. Thomas, S. Nieh"orster, S. Fabretti, N. Shepheard, O. Kuschel, K. K"upper, J. Wollschl"ager, P. Krzysteczko, and E. Chicca. In: *Frontiers in neuroscience* 9 (2015), p. 241.
- [68] J. L. Henning. In: *Computer* 33.7 (2000), pp. 28–35.
- [69] M. Davies, N. Srinivasa, T.-H. Lin, G. Chinya, Y. Cao, S. H. Choday, G. Dimou, P. Joshi, N. Imam, S. Jain, et al. In: *Ieee Micro* 38.1 (2018), pp. 82–99.
- [70] S. Xiang, J. Gong, Y. Zhang, X. Guo, Y. Han, A. Wen, and Y. Hao. In: *IEEE Journal of Quantum Electronics* 54.6 (2018), pp. 1–7.
- [71] S. B. Laughlin and T. J. Sejnowski. In: *Science* 301.5641 (2003), pp. 1870–1874.
- [72] E. Painkras, L. A. Plana, J. Garside, S. Temple, F. Galluppi, C. Patterson, D. R. Lester, A. D. Brown, and S. B. Furber. In: *IEEE Journal of Solid-State Circuits* 48.8 (2013), pp. 1943–1953.
- [73] L. El Soujji, A. Krishnan, R. Ravichandran, Y. Lee, M. On, X. Xiao, and S. Ben Yoo. In: *APL Photonics* 7.5 (2022).
- [74] A. Borji and L. Itti. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 113–120.
- [75] S. Gelly, L. Kocsis, M. Schoenauer, M. Sebag, D. Silver, C. Szepesv"ari, and O. Teytaud. In: *Communications of the ACM* 55.3 (2012), pp. 106–113.
- [76] C. Li, X. Zhang, J. Li, T. Fang, and X. Dong. In: *Photonix* 2.1 (2021), pp. 1–31.
- [77] J. MacMillan, E. E. Entin, and D. Serfaty. In: (2004).
- [78] D. Marković, A. Mizrahi, D. Querlioz, and J. Grollier. In: *Nature Reviews Physics* 2.9 (2020), pp. 499–510.
- [79] S. Park, S. Kim, B. Na, and S. Yoon. In: *2020 57th ACM/IEEE Design Automation Conference (DAC)*. IEEE. 2020, pp. 1–6.
- [80] C. D. Schuman, S. R. Young, J. P. Mitchell, J. T. Johnston, D. Rose, B. P. Maldonado, and B. C. Kaul. In: *2020 11th International Green and Sustainable Computing Workshops (IGSC)*. IEEE. 2020, pp. 1–8.
- [81] M. Hu, H. Li, Y. Chen, Q. Wu, G. S. Rose, and R. W. Linderman. In: *IEEE transactions on neural networks and learning systems* 25.10 (2014), pp. 1864–1878.
- [82] Y.-F. Lu, Y. Li, H. Li, T.-Q. Wan, X. Huang, Y.-H. He, and X. Miao. In: *IEEE Electron Device Letters* 41.8 (2020), pp. 1245–1248.
- [83] T. S. Lee and C. Choi. In: *Nanotechnology* 33.24 (2022), p. 245202.
- [84] M. Yan, N. Meisburger, T. Medini, and A. Shrivastava. In: *arXiv preprint arXiv:2201.12667* (2022).
- [85] H.-b. Chen and S. Fu. In: *2016 IEEE International Conference on Networking, Architecture and Storage (NAS)*. IEEE. 2016, pp. 1–4.
- [86] D. Liu, H. Yu, and Y. Chai. In: *Advanced Intelligent Systems* 3.2 (2021), p. 2000150.
- [87] F. A. Zwanenburg, A. S. Dzurak, A. Morello, M. Y. Simmons, L. C. Hollenberg, G. Klimeck, S. Rogge, S. N. Coppersmith, and M. A. Eriksson. In: *Reviews of modern physics* 85.3 (2013), p. 961.

- [88] L. Zhang, Z.-q. Liu, S.-W. Chen, Y.-d. Wang, W.-M. Long, Y.-h. Guo, S.-q. Wang, G. Ye, and W.-y. Liu. In: *Journal of Alloys and Compounds* 750 (2018), pp. 980–995.
- [89] F. Ponulak and A. Kasinski. In: *Acta neurobiologiae experimentalis* 71.4 (2011), pp. 409–433.
- [90] H. Altug, S.-H. Oh, S. A. Maier, and J. Homola. en. In: *Nature Nanotechnology* 17.11 (Jan. 2022), pp. 5–16. ISSN: 1748-3395.
- [91] K. Yao and Y. Zheng. eng. Springer series in optical sciences. Cham: Springer, 2023. ISBN: 978-3-031-20472-2.
- [92] T. Wang, S.-Y. Ma, L. G. Wright, T. Onodera, B. C. Richard, and P. L. McMahon. en. In: *Nature Communications* 13.11 (Jan. 2022), p. 123. ISSN: 2041-1723.
- [93] X. Xu, X. Xu, M. Tan, B. Corcoran, J. Wu, A. Boes, T. G. Nguyen, S. T. Chu, B. E. Little, D. Hicks, R. Morandotti, A. Mitchell, and D. J. Moss. In: *Nature* 589.7840 (2021). DOI: 10.1038/s41586-020-03063-0MAG ID: 3118265437PMID: 33408378, pp. 44–51.
- [94] B. Wu, H. Li, W. Tong, J. Dong, and X. Zhang. In: *Optical Materials Express* (Feb. 2022). DOI: 10.1364/ome.447330MAG ID: 4210275959.
- [95] R. Amin, R. Maiti, C. Carfano, Z. Ma, M. H. Tahersima, Y. Lilach, D. Ratnayake, H. Dalir, and V. J. Sorger. In: *Apl Photonics* 3.12 (2018).
- [96] T. Zhou, M. Tang, G. Xiang, B. Xiang, S. Hark, M. Martin, T. Baron, S. Pan, J.-S. Park, Z. Liu, et al. In: *Nature communications* 11.1 (2020), p. 977.
- [97] Q. Xu, B. Schmidt, S. Pradhan, and M. Lipson. In: *nature* 435.7040 (2005), pp. 325–327.
- [98] F. Xia, T. Mueller, Y.-m. Lin, A. Valdes-Garcia, and P. Avouris. In: *Nature nanotechnology* 4.12 (2009), pp. 839–843.
- [99] M. Cazzanelli, F. Bianco, E. Borga, G. Pucker, M. Ghulinyan, E. Degoli, E. Luppi, V. Véniard, S. Ossicini, D. Modotto, et al. In: *Nature materials* 11.2 (2012), pp. 148–154.
- [100] R. Li, X. Gu, K. Li, Y. Huang, Z. Li, and Z. Zhang. In: *Optical Materials Express* 11.7 (2021), pp. 2122–2133.
- [101] R. Li, C. Zhang, W. Xie, Y. Gong, F. Ding, H. Dai, Z. Chen, F. Yin, and Z. Zhang. In: *Nanophotonics* 12.2 (2023), pp. 319–334.
- [102] M. A. Iqbal, N. Ashraf, W. Shahid, M. Awais, A. K. Durrani, K. Shahzad, and M. Ikram. In: (2021).
- [103] R. Li, C. Zhang, S. Mao, H. Huang, M. Zhong, Y. Cui, X. Zhou, F. Yin, S. Theodoridis, and Z. Zhang. 14 pages, 9 graphics. Aug. 2023.
- [104] K. Hirose, Y. Liang, Y. Kurosaka, A. Watanabe, T. Sugiyama, and S. Noda. In: *Nature photonics* 8.5 (2014), pp. 406–411.
- [105] M. Kneissl, A. Knorr, S. Reitzenstein, and A. Hoffmann. Vol. 194. Springer, 2020.
- [106] M. Ohtsu, K. Kobayashi, T. Kawazoe, S. Sangu, and T. Yatsui. In: *IEEE Journal of Selected Topics in Quantum Electronics* 8.4 (2002), pp. 839–862.
- [107] P. N. Prasad. John Wiley & Sons, 2004.
- [108] F. Monticone and A. Alu. In: *Reports on Progress in Physics* 80.3 (2017), p. 036401.
- [109] A. Karabchevsky, A. Katiyi, A. S. Ang, and A. Hazan. In: *Nanophotonics* 9.12 (2020), pp. 3733–3753.
- [110] M. C. Roco, M. C. Hersam, C. A. Mirkin, E. L. Hu, M. Brongersma, and A. Baca. In: *Nanotechnology Research Directions for Societal Needs in 2020: Retrospective and Outlook* (2011), pp. 417–444.
- [111] N. Xi and K. Lai. William Andrew, 2011.
- [112] M. Ohtsu, T. Kawazoe, T. Yatsui, and M. Naruse. In: *IEEE Journal of Selected Topics in Quantum Electronics* 14.6 (2008), pp. 1404–1417.
- [113] D. Dai. In: *Journal of Lightwave Technology* 35.4 (2016), pp. 572–587.
- [114] P. B. Deotare, M. W. McCutcheon, I. W. Frank, M. Khan, and M. Lončar. In: *Applied Physics Letters* 94.12 (2009).
- [115] J. Ma, T. Zhou, M. Tang, H. Li, Z. Zhang, X. Xi, M. Martin, T. Baron, H. Liu, Z. Zhang, S. Chen, and X. Sun. In: arXiv:2106.13838 (June 2021). arXiv:2106.13838 [cond-mat, physics:physics].
- [116] B. Xie, G. Su, H.-F. Wang, F. Liu, L. Hu, S.-Y. Yu, P. Zhan, M.-H. Lu, Z. Wang, and Y.-F. Chen. In: *Nature communications* 11.1 (2020), p. 3768.
- [117] M. S. Alias, M. Tangi, J. A. Holguin-Lerma, E. Stegenburgs, A. A. Alatawi, I. Ashry, R. C. Subedi, D. Priante, M. K. Shakfa, T. K. Ng, et al. In: *Journal of Nanophotonics* 12.4 (2018), pp. 043508–043508.
- [118] B. Momeni, S. Yegnanarayanan, M. Soltani, A. A. Eftekhar, E. S. Hosseini, and A. Adibi. In: *Journal of Nanophotonics* 3.1 (2009), p. 031001.
- [119] L. Xu, T. F. De Lima, H.-T. Peng, S. Bilodeau, A. Tait, B. J. Shastri, and P. R. Prucnal. In: *IEEE Journal of Selected Topics in Quantum Electronics* 28.6: High Density Integr. Multipurpose Photon. Circ. (2022), pp. 1–9.

- [120] M. S. Nezami, T. F. De Lima, M. Mitchell, S. Yu, J. Wang, S. Bilodeau, W. Zhang, M. Al-Qadasi, I. Taghavi, A. Tofini, et al. In: *IEEE Journal of Selected Topics in Quantum Electronics* 29.2: Optical Computing (2022), pp. 1–11.
- [121] J. Hansryd, P. A. Andrekson, M. Westlund, J. Li, and P.-O. Hedekvist. In: *IEEE Journal of Selected Topics in Quantum Electronics* 8.3 (2002), pp. 506–520.
- [122] A. K. Kodi and A. Louri. In: *IEEE journal of selected topics in Quantum Electronics* 17.2 (2010), pp. 384–395.
- [123] S. Carrillo, J. Harkin, L. J. McDaid, F. Morgan, S. Pande, S. Cawley, and B. McGinley. In: *IEEE Transactions on Parallel and Distributed Systems* 24.12 (2012), pp. 2451–2461.
- [124] M. Moralis-Pegios, G. Mourgas-Alexandris, A. Tsakyridis, G. Giampougiannis, A. Totovic, G. Dabos, and N. Pleros. In: *Optical Interconnects XXII*. Vol. 12007. SPIE. 2022, pp. 25–30.
- [125] G. Dabos, D. V. Bellas, R. Stabile, M. Moralis-Pegios, G. Giampougiannis, A. Tsakyridis, A. Totovic, E. Lidorikis, and N. Pleros. In: *Optical Materials Express* 12.6 (2022), pp. 2343–2367.
- [126] M. A. Nahmias, B. J. Shastri, A. N. Tait, T. F. De Lima, and P. R. Prucnal. In: *Optics and Photonics News* 29.1 (2018), pp. 34–41.
- [127] R. Dangel, A. La Porta, D. Jubin, F. Horst, N. Meier, M. Seifried, and B. J. Offrein. In: *IEEE Journal of selected topics in quantum electronics* 24.4 (2018), pp. 1–11.
- [128] X. Xiao, R. Proietti, G. Liu, H. Lu, P. Fotouhi, S. Werner, Y. Zhang, and S. B. Yoo. In: *IEEE Journal of Selected Topics in Quantum Electronics* 26.2 (2019), pp. 1–10.
- [129] Y. Zhang, A. Samanta, K. Shang, and S. B. Yoo. In: *IEEE Journal of Selected Topics in Quantum Electronics* 26.2 (2020), pp. 1–10.
- [130] X. Guo, J. Xiang, Y. Zhang, and Y. Su. In: *Advanced Photonics Research* 2.6 (2021), p. 2000212.
- [131] J. M. Shainline, S. M. Buckley, R. P. Mirin, and S. W. Nam. In: *Physical Review Applied* 7.3 (2017), p. 034013.
- [132] F. Catthoor, S. Mitra, A. Das, and S. Schaafsma. In: *CMOS Circuits for Biological Sensing and Processing* (2018), pp. 315–340.
- [133] S. Jiao, J. Liu, L. Zhang, F. Yu, G. Zuo, J. Zhang, F. Zhao, W. Lin, and L. Shao. en. In: *Opto-Electronic Science* 1.9 (Sept. 2022), pp. 220010–22. ISSN: 2097-0382.
- [134] H. Hu, X. Zhang, and S. Zhao. In: *Cogent Physics* 4.1 (Jan. 2017). Ed. by L. Zhang, p. 1388156. ISSN: null.
- [135] J.-Y. Kim, J.-M. Kang, T.-Y. Kim, and S.-K. Han. EN. In: *Journal of Lightwave Technology* 24.9 (Sept. 2006), p. 3392.
- [136] C. Liu, Q. Ma, Z. J. Luo, Q. R. Hong, Q. Xiao, H. C. Zhang, L. Miao, W. M. Yu, Q. Cheng, L. Li, and T. J. Cui. en. In: *Nature Electronics* 5.22 (Feb. 2022), pp. 113–122. ISSN: 2520-1131.
- [137] X. Luo, Y. Hu, X. Ou, X. Li, J. Lai, N. Liu, X. Cheng, A. Pan, and H. Duan. en. In: *Light: Science & Applications* 11.11 (May 2022), p. 158. ISSN: 2047-7538.
- [138] D. T. Tan. In: *Advanced Photonics Research* 2.9 (2021), p. 2100010.
- [139] G.-J. Tang, X.-T. He, F.-L. Shi, J.-W. Liu, X.-D. Chen, and J.-W. Dong. en. In: *Laser & Photonics Reviews* 16.4 (2022), p. 2100300. ISSN: 1863-8899.
- [140] H. Wang, L. Sun, Y. He, G. Tang, S. An, Z. Wang, Y. Du, Y. Zhang, L. Yuan, X. He, J. Dong, and Y. Su. en. In: *Laser & Photonics Reviews* 16.6 (2022), p. 2100631. ISSN: 1863-8899.
- [141] H. Wang, G. Tang, Y. He, Z. Wang, X. Li, L. Sun, Y. Zhang, L. Yuan, J. Dong, and Y. Su. en. In: *Light: Science & Applications* 11.11 (Oct. 2022), p. 292. ISSN: 2047-7538.
- [142] M. M. Plakhotnyuk, M. Gaudig, R. S. Davidsen, J. M. Lindhard, J. Hirsch, D. Lausch, M. S. Schmidt, E. Stamate, and O. Hansen. In: *Journal of Applied Physics* 122.14 (2017).
- [143] X. Yu, T. Zhang, and Y. Li. In: *Polymers* 12.8 (2020), p. 1637.
- [144] M. D. Levenson, N. Viswanathan, and R. A. Simpson. In: *IEEE Transactions on electron devices* 29.12 (1982), pp. 1828–1836.
- [145] A. Kaganskiy, M. Gschrey, A. Schleehahn, R. Schmidt, J.-H. Schulze, T. Heindel, A. Strittmatter, S. Rodt, and S. Reitzenstein. In: *Review of Scientific Instruments* 86.7 (2015).
- [146] M. Gschrey, R. Schmidt, J.-H. Schulze, A. Strittmatter, S. Rodt, and S. Reitzenstein. In: *Journal of Vacuum Science & Technology B* 33.2 (2015).
- [147] J. Viheriälä, T. Niemi, J. Kontio, and M. Pessa. In: *Recent Optical and Photonic Technologies* (2010), pp. 275–298.
- [148] H. Lan. In: *Updates in Advanced Lithography*. Ed. by S. Hosaka. Rijeka: IntechOpen, 2013. Chap. 7.
- [149] S. H. Teo, A. Liu, G. Sia, C. Lu, J. Singh, and M. Yu. In: *International Journal of Nanoscience* 4.04 (2005), pp. 567–574.
- [150] Y. Fainman, M. Nezhad, D. Tan, K. Ikeda, O. Bondarenko, and A. Grieco. In: *Applied Optics* 52.4 (2013), pp. 613–624.

- [151] W. Freude, C. Poulton, C. Koos, J. Brosi, F. Glockler, J. Wang, G.-A. Chakam, and M. Fujii. In: *Proceedings of 2004 6th International Conference on Transparent Optical Networks (IEEE Cat. No. 04EX804)*. Vol. 1. IEEE. 2004, pp. 4–9.
- [152] S. K. Selvaraja, W. Bogaerts, and D. Van Thourhout. In: *Optics Communications* 284.8 (2011), pp. 2141–2144.
- [153] T. Campbell, C. Williams, O. Ivanova, and B. Garrett. In: *Technologies, Potential, and Implications of Additive Manufacturing, Atlantic Council, Washington, DC 3* (2011), pp. 1–16.
- [154] J. E. Melzer and E. McLeod. In: *Nanophotonics* 9.6 (2020), pp. 1373–1390.
- [155] C. Roques-Carmes, Z. Lin, R. E. Christiansen, Y. Salamin, S. E. Kooi, J. D. Joannopoulos, S. G. Johnson, and M. Soljacic. In: *ACS Photonics* 9.1 (2022), pp. 43–51.
- [156] K. B. Fritzler and V. Y. Prinz. In: *Physics-Uspekhi* 62.1 (2019), p. 54.
- [157] H. Y. Jeong, E. Lee, S.-C. An, Y. Lim, and Y. C. Jun. In: *Nanophotonics* 9.5 (2020), pp. 1139–1160.
- [158] J. Pyo, J. T. Kim, J. Lee, J. Yoo, and J. H. Je. In: *Advanced Optical Materials* 4.8 (2016), pp. 1190–1195.
- [159] S. Manipatruni, D. E. Nikonorov, C.-C. Lin, T. A. Gosavi, H. Liu, B. Prasad, Y.-L. Huang, E. Bonturim, R. Ramesh, and I. A. Young. In: *Nature* 565.7737 (2019), pp. 35–42.
- [160] A. Ryou, J. Whitehead, M. Zhelyeznyakov, P. Anderson, C. Keskin, M. Bajcsy, and A. Majumdar. In: *Photonics Research* 9.4 (2021), B128–B134.
- [161] T. Wang, M. M. Sohoni, L. G. Wright, M. M. Stein, S.-Y. Ma, T. Onodera, M. G. Anderson, and P. L. McMahon. In: *Nature Photonics* 17.5 (2023), pp. 408–415.
- [162] Z. Wang, Y. Chong, J. D. Joannopoulos, and M. Soljačić. en. In: *Nature* 461.72657265 (Oct. 2009), pp. 772–775. ISSN: 1476-4687.
- [163] T. Zhou, J. Ma, M. Tang, H. Li, M. Martin, T. Baron, H. Liu, S. Chen, X. Sun, and Z. Zhang. In: *ACS PHOTONICS* 9.12 (Dec. 2022), pp. 3824–3830. ISSN: 2330-4022.
- [164] B. Bahari, A. Ndao, F. Vallini, A. El Amili, Y. Fainman, and B. Kanté. In: *Science* 358.6363 (Nov. 2017), pp. 636–640.
- [165] Z.-K. Shao, H.-Z. Chen, S. Wang, X.-R. Mao, Z.-Q. Yang, S.-L. Wang, X.-X. Wang, X. Hu, and R.-M. Ma. en. In: *Nature Nanotechnology* 15.11 (Jan. 2020), pp. 67–72. ISSN: 1748-3395.
- [166] W. Zhang, X. Xie, H. Hao, J. Dang, S. Xiao, S. Shi, H. Ni, Z. Niu, C. Wang, K. Jin, X. Zhang, and X. Xu. en. In: *Light: Science & Applications* 9.11 (June 2020), p. 109. ISSN: 2047-7538.
- [167] L. Yang, G. Li, X. Gao, and L. Lu. en. In: *Nature Photonics* 16.44 (Apr. 2022), pp. 279–283. ISSN: 1749-4893.
- [168] S. Han, Y. Chua, Y. Zeng, B. Zhu, C. Wang, B. Qiang, Y. Jin, Q. Wang, L. Li, A. G. Davies, E. H. Linfield, Y. Chong, B. Zhang, and Q. J. Wang. en. In: *Nature Communications* 14.1 (Feb. 2023), p. 707. ISSN: 2041-1723.
- [169] Y. Zeng, U. Chattopadhyay, B. Zhu, B. Qiang, J. Li, Y. Jin, L. Li, A. G. Davies, E. H. Linfield, B. Zhang, Y. Chong, and Q. J. Wang. en. In: *Nature* 578.77947794 (Feb. 2020), pp. 246–250. ISSN: 1476-4687.
- [170] M. I. Shalaev, W. Walasik, A. Tsukernik, Y. Xu, and N. M. Litchinitser. en. In: *Nature Nanotechnology* 14.11 (Jan. 2019), pp. 31–34. ISSN: 1748-3395.
- [171] G. Arregui, J. Gomis-Bresco, C. M. Sotomayor-Torres, and P. D. Garcia. en. In: *Physical Review Letters* 126.2 (Jan. 2021), p. 027403. ISSN: 0031-9007, 1079-7114.
- [172] L. He, H. Zhang, W. Zhang, Y. Wang, and X. Zhang. en. In: *New Journal of Physics* 23.9 (Sept. 2021), p. 093026. ISSN: 1367-2630.
- [173] X.-T. He, E.-T. Liang, J.-J. Yuan, H.-Y. Qiu, X.-D. Chen, F.-L. Zhao, and J.-W. Dong. en. In: *Nature Communications* 10.11 (Feb. 2019), p. 872. ISSN: 2041-1723.
- [174] M. Yoshida, M. De Zoysa, K. Ishizaki, Y. Tanaka, M. Kawasaki, R. Hatsuda, B. Song, J. Gelleta, and S. Noda. In: *Nature materials* 18.2 (2019), pp. 121–128.
- [175] S. Noda, K. Kitamura, T. Okino, D. Yasuda, and Y. Tanaka. In: *IEEE Journal of Selected Topics in Quantum Electronics* 23.6 (2017), pp. 1–7.
- [176] Q. Quan, P. B. Deotare, and M. Loncar. In: *Applied Physics Letters* 96.20 (2010), p. 203102.
- [177] C. J. Chang-Hasnain. In: *IEEE Journal of Selected Topics in Quantum Electronics* 6.6 (2000), pp. 978–987.
- [178] S. M. Sze, Y. Li, and K. K. Ng. John wiley & sons, 2021.
- [179] J. Lee, I. Karnadi, J. T. Kim, Y.-H. Lee, and M.-K. Kim. In: *ACS Photonics* 4.9 (2017), pp. 2117–2123.
- [180] L. Huang, Q. A. Tanguy, J. E. Fröch, S. Mukherjee, K. F. Böhringer, and A. Majumdar. In: *Nanophotonics* 13.7 (2024), pp. 1191–1196.
- [181] S. Srivastava and A. Acharya. In: *Device Circuit Co-Design Issues in FETs* (2023), p. 231.
- [182] C. Errando-Herranz, A. Y. Takabayashi, P. Edinger, H. Sattari, K. B. Gylfason, and N. Quack. In: *IEEE Journal of Selected Topics in Quantum Electronics* 26.2 (2019), pp. 1–16.
- [183] J. M. P. Fernández. PhD thesis. Universidad Carlos III de Madrid, 2022.

- [184] W. Zhou. In: (2013).
- [185] J. Hu, S. Bandyopadhyay, Y.-h. Liu, and L.-y. Shao. In: *Frontiers in Physics* 8 (2021), p. 586087.
- [186] D. Zhao, Z. Lin, W. Zhu, H. J. Lezec, T. Xu, A. Agrawal, C. Zhang, and K. Huang. In: *Nanophotonics* 10.9 (2021), pp. 2283–2308.
- [187] M. J. Filipovich, Z. Guo, M. Al-Qadasi, B. A. Marquez, H. D. Morison, V. J. Sorger, P. R. Prucnal, S. Shekhar, and B. J. Shastri. In: *Optica* 9.12 (2022), pp. 1323–1332.
- [188] T. Wu, M. Menarini, Z. Gao, and L. Feng. In: *Nature Photonics* 17.8 (2023), pp. 710–716.
- [189] S. Pai, Z. Sun, T. W. Hughes, T. Park, B. Bartlett, I. A. Williamson, M. Minkov, M. Milanizadeh, N. Abebe, F. Morichetti, et al. In: *Science* 380.6643 (2023), pp. 398–404.
- [190] E. Pelucchi, G. Fagas, I. Aharonovich, D. Englund, E. Figueroa, Q. Gong, H. Hannes, J. Liu, C.-Y. Lu, N. Matsuda, et al. In: *Nature Reviews Physics* 4.3 (2022), pp. 194–208.
- [191] A. Laucht, F. Hohls, N. Ubbelohde, M. F. Gonzalez-Zalba, D. J. Reilly, S. Stobbe, T. Schröder, P. Scarlino, J. V. Koski, A. Dzurak, et al. In: *Nanotechnology* 32.16 (2021), p. 162003.
- [192] P. Khanpara and S. Tanwar. In: *A Roadmap to Industry 4.0: Smart Production, Sharp Business and Sustainable Development* (2020), pp. 171–185.
- [193] M. Mehrpouya, A. Dehghanhadikolaei, B. Fotovvati, A. Vosooghnia, S. S. Emamian, and A. Gisario. In: *Applied Sciences* 9.18 (2019), p. 3865.
- [194] K. Dhananjay, P. Shukla, V. F. Pavlidis, A. Coskun, and E. Salman. In: *IEEE Transactions on Circuits and Systems II: Express Briefs* 68.3 (2021), pp. 837–843.
- [195] S. Zhang, Z. Li, H. Zhou, R. Li, S. Wang, K.-W. Paik, and P. He. In: *e-Prime-Advances in Electrical Engineering, Electronics and Energy* (2022), p. 100052.
- [196] Z. Chen, J. Zhang, S. Wang, and C.-P. Wong. In: *Fundamental Research* (2023).
- [197] P. Chaourani. PhD thesis. KTH Royal Institute of Technology, 2019.
- [198] P. Girard, Y. Cheng, A. Virazel, W. Zhao, R. Bishnoi, and M. B. Tahoori. In: *Proceedings of the IEEE* 109.2 (2020), pp. 149–169.
- [199] W. Chen, B. Wu, and H. Wang. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 9727–9737.



Renjie Li received his B.S. in Aerospace and Mechanical Engineering from University of California, Los Angeles, CA, USA. In August 2020, he joined the Shenzhen Key Laboratory of Semiconductor Lasers and The Chinese University of Hong Kong, Shenzhen to pursue a doctoral degree. His main research interest is the design and optimization of semiconductor laser devices, machine learning, derivative-free optimization, and LLM.



Yuanhao Gong received his B.Eng. from Northwestern Polytechnical University in Material Science and Engineering, Xi'an, China. In July 2021, he joined the Nano Opto-Electronics Laboratory in the Chinese University of Hong Kong, Shenzhen as a PhD student. His main research interests include the fabrication and growth of micro-cavity lasers on silicon, wafer epitaxy, silicon-on-insulator growth, and photonic crystal surface emitting lasers.



Hai Huang received his B.Eng. from Harbin Institute of Technology, Harbin, China. Now he is pursuing his PhD at The Chinese University of Hong Kong, Shenzhen. His current research interests mainly involve the design, simulation, and fabrication of PCSEL, TCSEL and other micro-nano laser devices.



Yuze Zhou received his B.Eng. in Department of Materials Science and Engineering from Chongqing University, Chongqing, China. In July 2023, he joined the Shenzhen Key Laboratory of Semiconductor Lasers and The Chinese University of Hong Kong, Shenzhen. His main research interest is the design, characterization, and fabrication of semiconductor laser devices.



Sixuan Mao is an undergraduate student in the School of Science and Engineering of the Chinese University of Hongkong (Shenzhen), Shenzhen, China. In June 2023, he joined the Shenzhen Key Laboratory of Semiconductor Lasers. His research interests include nanophotonic devices, LLM, and deep learning.



Prof. Zhaoyu Zhang received his B.S. and M.S. degrees in Applied Mechanics from University of Science and Technology of China, Hefei, China, in 1998 and 2001 respectively. He received Ph.D. degree from California Institute of Technology, Pasadena USA in 2007 in Electrical Engineering. From 2008 to 2011, he worked at University of California, Berkeley as a postdoctoral fellow in the College of Chemistry, with a joint appointment with Lawrence Berkeley National Laboratory. From 2011 to 2015, he worked in Peking University as an Associate Professor and led the team of “Nano-OptoElectronics Lab (NOEL)”. In 2015, he and his team moved to The Chinese University of Hong Kong, Shenzhen. In 2016, he was approved to set up the Shenzhen Key Laboratory of Semiconductor lasers and be the director. His main achievements include the first demonstration of red-emission photonic crystal lasers, wavelength-scale micro-lasers with physical size smaller than 1 micron, microfluidic microlasers based on dye materials, as well as the first demonstration of photonic crystal lasers directly grown on silicon substrates. He has published more than 30 peer-reviewed papers on renowned journals including Nature Communications, Light: Science and Applications, ACS Photonics, Advanced Materials, Nanophotonics, Physics Review Letters, Optica, Photonics research, Optics Letters, Applied Physics Letters, etc.